1999 Named Entity Recognition Task Definition

Version 1.4

August 27, 1999

Nancy Chinchor (chinchor@gso.saic.com) (1)

Erica Brown (erica@gso.saic.com) (2)

Lisa Ferro (lferro@mitre.org) (3)

Patty Robinson (4)

Acknowledgments

This work was supported by DARPA under contract DAAB07-99-C-C201 and is subject to the 'Rights in Data' clause at DFARS at 252.227-7014.

This work was performed as part of a collaborative effort between MITRE and SAIC.

Authors (1) and (2) are with
Science Applications International Corporation
10260 Campus Pt. Dr. M/S X2
San Diego, California 92121

Author (3) is with (and author (4) was formerly with)
The MITRE Corporation
202 Burlington Rd.
Bedford, Massachusetts 01730

Contents

1 INTRODUCTION 1
1.1 Scope 1
1.2 Performance Evaluation 1
1.3 Format of Examples in this Document 1
2 TASK OVERVIEW 1
2.1 Markup Description 2
2.2 Authoritative Reference Materials 2
2.3 Named Entities (ENAMEX tag element) 2
2.4 Temporal Expressions (TIMEX tag element) 2
2.5 Number Expressions (NUMEX tag element) 2
3 NOTATION RESERVED FOR USE IN THE ANSWER KEYS 3
3.1 Expressing Alternative Attribute Values 3
3.2 Expressing Alternative String Boundaries (ALT Attribute) 3
3.3 Expressing Optional Markup (STATUS Attribute) 3
4 GUIDELINES FOR MARKUP OF EXCEPTIONAL CONSTRUCTIONS 3
4.1 Conjunction and Elision in Multi-modifier Expressions and Multi-Name/Number Expressions 3
4.1.1 Multi-modifier Expressions 3
4.1.2 Multi-name (or Multi-number) Expressions 3
4.2 Numeric Range Expressions 4
4.3 Effects of Tokenization Conventions 4
4.4 Nested Expressions 5
5 ENAMEX: SPECIFIC GUIDELINES 5
5.1 Guidelines That Pertain to All TYPEs (PERSON, LOCATION, ORGANIZATION) 5
5.1.1 Entity Expressions that Modify Non-taggables 5
5.1.2 Entity-Strings Embedded in Entity Expressions 6
5.1.3 Entity Expressions that Possess Other Entity Expressions 6
5.1.4 Entity Expression Aliases 7
5.1.4.1 Taggable Aliases 7
5.1.4.2 Non-taggable Aliases 8
5.1.5 Quotation Marks Around an Entity Name 8
5.1.6 The Definite Article in an Entity Name 9
5.1.7 Non-decomposable Names 9
5.1.8 Miscellaneous Non-taggables 10
5.1.8.1 Figures of Speech 10
5.1.8.2 Non-taggable Proper Names 10
5.2 Guidelines That Pertain Only to PERSON 10
5.2.1 Titles of PERSONs 10
5.2.1.1 Titles vs. Generational Designators 10
5.2.1.2 Entities that Modify Persons/Titles 11
5.2.2 Family Entity Expressions 11
5.2.3 Names of Animals 11

5.2.4 Saints and other Religious Figures 11
5.2.5 Fictional Characters 12
5.2.6 Fictional Animals and Non-human Characters 12
5.2.7 Miscellaneous Personal Non-taggables 12
5.3 Guidelines That Pertain Only to LOCATION 13
5.3.1 Possible Embedded Locative Entity-Strings 14
5.3.2 Locative Entity Expressions Tagged in Succession 14
5.3.3 Locative Designators and Specifiers 15
5.3.3.1 Locative Non-taggables: The Postposed Partitive Specifier 15
5.3.3.2 Exceptional Locative Specifiers Used as Entity Expressions 15
5.3.4 Transnational and Subnational Region Names 15
5.3.4.1 Transnational Locative Entity Expressions 15
5.3.4.2 Subnational Region Names 16
5.3.5 Time Modifiers of Locative Entity Expressions 16
5.3.6 Space Modifiers of Locative Entity Expressions 16
5.3.7 Miscellaneous Locative Non-taggables 17
5.3.7.1 Adjectival Forms of Location Names 17
5.4 Guidelines That Pertain Only to ORGANIZATION 17
5.4.1 Corporate or Organization Designators 17
5.4.2 Miscellaneous ORGANIZATION-type Entity Expressions 17
5.4.2.1 Broadcasting Stations 18
5.4.2.2 Legislative Bodies 18
5.4.2.3 Event Organizers 18
5.4.2.4 Membership Affiliation 18
5.4.2.5 ORGANIZATION-related Facilities 19
5.4.3 Decomposable Product Names 19
5.4.4 Metonymy 20
5.4.5 Generic ORGANIZATION-like Non-taggables 20
6 TIMEX: SPECIFIC GUIDELINES 21
6.1 Introduction 21
6.2 Absolute Temporal Expressions - TIME & DATE 21
6.3 Scope of Temporal Expressions 22
6.4 Temporal Expressions Containing Adjacent Absolute and Relative Strings 23
6.5 Holidays 23
6.6 Locative Entity-Strings Embedded in Temporal Expressions 23
6.7 Temporal Expressions Based on Alternate Calendars 24
6.8 DURATION 24
6.8.1 Scope of DURATION Expressions 25
6.8.2 Non-taggable Durations 25
6.8.3 Distinguishing DURATION from DATE/TIME 25
6.8.4 Distinguishing TIMEX DURATION from NUMEX MEASURE 25
7 NUMEX: SPECIFIC GUIDELINES 25
7.1 Scope of Numeric Expressions 26
7.1.1 Numeric Expressions Plus Units of Measure 26
7.1.2 Numeric Expressions without Units of Measure 27
7.2 Numeric Expressions Appearing in Succession 27

7.3 Approximators and Multipliers in the Modification of Numeric Expressions 27
7.3.1 Approximators 27
7.3.2 Multipliers 28
7.3.3 Orders of Magnitude 28
7.3.4 The Indefinite Article 28
7.4 TYPE-Specific Guidelines 29
7.4.1 MONEY 29
7.4.2 MEASURE 29
7.4.2.1 Standard Measurement Units 29
7.4.2.2 Measurement Phrases 30
7.4.2.3 Some Exceptional Phrases Tagged as MEASURE 32
7.4.2.4 Combinations of Measured Quantities 32
7.4.2.5 Numeric Expressions Not Tagged as MEASURE 33
7.4.3 PERCENT 33
7.4.4 CARDINAL 33
7.5 Decomposable Idioms 34
7.6 Non-taggagable Numbers 34
7.6.1 Non-taggable Numbers within Names 34
7.6.2 Non-taggable Numbers within Temporal Expressions 35
7.6.3 Non-taggable Pronoun "One" 35
7.6.4 Non-taggable Ordinals 35
7.7 Non-taggable Non-Numeric Expressions 36
APPENDIX A. TOKENIZATION RULES 37
APPENDIX B. SPECIAL FEATURES FOUND ONLY IN TRANSCRIBED SPEECH ... 41
APPENDIX C. GUIDELINES FOR MARKUP OF NEWSWIRE DATA 45

1 INTRODUCTION

1.1 Scope

The Named Entity task consists of three subtasks: entity names, temporal expressions, and number expressions. The expressions to be annotated are “unique identifiers” of entities (persons, locations, organizations), times (dates, times, and durations), and quantities (money, measures, percents, and cardinal numbers).

For many text processing systems, such identifiers are recognized primarily using local pattern-matching techniques. The TEI (Text Encoding Initiative) Guidelines for Electronic Text Encoding and Interchange cover such identifiers (plus abbreviations) together in section 6.4 of the TEI Guidelines and explain that the identifiers comprise “textual features which it is often convenient to distinguish from their surrounding text. Names, dates and numbers are likely to be of particular importance to the scholar treating a text as source for a database; distinguishing such items from the surrounding text is however equally important to the scholar primarily interested in lexis.”

The task is to identify all instances of the three types of expressions in each text in the test set and to subcategorize the expressions. The original texts contain some SGML tags already; the Named Entity task is to be performed on all of the text, except as specified in Appendix C.

The system must produce a single, unambiguous output for any relevant string in the text; thus, this evaluation is not based on a view of a pipelined system architecture in which Named Entity recognition would be completely handled as a preprocess to sentence and discourse analysis. The task requires that the system recognize what a string represents, not just its superficial appearance. Sometimes, the right answer is superficially apparent, as in the case of many numerical expressions, and can be obtained by local pattern-matching techniques. In other cases, the right answer is not superficially apparent, as when a single capitalized word could represent the name of a location, person, or organization, and the answer may have to be obtained using techniques that draw information from a larger context or from reference lists. Transcriptions of speech lack most capitalization and punctuation found in electronic newswire articles; this missing information makes certain decisions regarding proper names more difficult. Speech recognizers may have trouble identifying numbers accurately, which comprise a large portion of temporal and numerical expressions.

The three subtasks correspond to three SGML tag elements: ENAMEX, TIMEX, and NUMEX. The subcategorization is captured by a SGML tag attribute called TYPE, which is defined to have a different set of possible values for each tag element. The markup is described in section 2, below.

1.2 Performance Evaluation

Scoring of this task will be done using an error-based metric proposed at the February 1999 HUB-4 meeting by John Makoul of BBN as well as the same kinds of metrics that are used for scoring template-filling (information extraction) tasks. For specific information on the scoring, refer to “NE99 Scoring System User’s Manual,” prepared for NE99 by SAIC.

Cumulative scores will be generated at several levels of description of the task, e.g.,

- across subtasks,
- for each subtask,
- for the subcategorization aspect of each subtask.

1.3 Format of Examples in this Document

Examples in this document encompass both text (originally written) and speech (transcribed by humans). To distinguish between the sources of these examples, two different fonts have been used for the source material. In both cases, the resultant “marked-up” text is given in a third font.

> “Original text examples are presented in this font.”
> "Original speech examples are presented in this font."
> `Markup (with appropriate tags included) is presented in this font.`

Sections using speech examples will follow the rules for speech markup (see appendices), while text examples will follow the rules for text markup.

2 TASK OVERVIEW

2.1 Markup Description

The output of the systems to be evaluated will be in the form of SGML text markup. The only insertions allowed during tagging are tags enclosed in angled brackets. No extra whitespace or carriage returns are to be inserted; otherwise, the offset count would change, which would adversely affect scoring for newswire, but have little or no effect for transcriptions.

The markup will have the following form:

```
<B_ELEMENT-NAME ATTR-NAME="ATTR-VALUE" ...>text-string<E_ELEMENT-NAME>
```

Example:

```
<B_ENAMEX TYPE="ORGANIZATION">Taga Co.<E_ENAMEX>
```

The markup is defined in SGML Document Type Descriptions (DTDs), written for NE99 use and maintained by personnel at SAIC. The DTDs enable annotators and system developers to use SGML validation tools to check the correctness of the SGML-tagged texts produced by the annotator or the system. The validation tools are available to NE99 participants in the file called sgml-tools and in the form of the scorer’s parser both available via anonymous ftp from ftp.muc.saic.com (or online.muc.saic.com) in the SPEECH subdirectory.

Annotators are using the Alembic Workbench, a software tool provided for NE99 by MITRE to assist in generating the answer keys to be used for system development, training, and testing.

2.2 Authoritative Reference Materials

In order to encourage consistency and reduce ambiguity regarding taggability, four reference works have been chosen which are currently available. These resources are:

- *The Chicago Manual of Style*, (Chicago: University of Chicago Press, 1993), fourteenth edition;
- Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., *A Grammar of Contemporary English*, (London: Longman Group Limited, 1984);
- *Merriam-Webster’s Geographical Dictionary*, (Springfield, MA: Merriam-Webster, Incorporated, 1997), third edition.
- *The American Heritage Dictionary of the English Language*, (Boston: Houghton Mifflin, 1992), third edition.

These are to be used when context and world knowledge are not sufficient to decide upon the most accurate markup.

2.3 Named Entities (ENAMEX tag element)

This subtask is limited to proper names, acronyms, and other unique identifiers, which are categorized via the TYPE attribute as follows:

PERSON: named person, family, or certain designated non-human individuals
ORGANIZATION: named corporate, governmental, or other organizational entity
LOCATION: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.) and astronomical locations.

2.4 Temporal Expressions (TIMEX tag element)

This subtask is for “absolute” temporal expressions only. The tagged tokens are categorized via the TYPE attribute as follows:

DATE: complete or partial date expression
TIME: complete or partial expression of time of day
DURATION: a measurement of time elapsed or period of time during which something lasts

2.5 Number Expressions (NUMEX tag element)

This subtask is for useful numeric expressions, including numbers with their units of measure and numbers in other quantifying positions. The tagged tokens are categorized via the TYPE attribute as follows:

MONEY: monetary expression
MEASURE: standard numeric measurement phrases such as age, area, distance, energy, speed, temperature, volume, and weight, plus syntactically-defined measurement phrases
PERCENT: percentage (a fraction expressed in terms of hundredths)
CARDINAL: a numerical count or quantity of some object (in the form of whole numbers, decimals, or fractions)

The numbers may be expressed in either numeric or alphabetic form. Note that in most speech transcriptions, numbers and expressions or symbols will be spelled out.

3 NOTATION RESERVED FOR USE IN THE ANSWER KEYS

Premodifiers listed in the scoring configuration file (currently, “a,” “an,” and “the”) can be included unconsciously during human markup, and systems should not be penalized. The answer keys are not consistent with respect to including or excluding articles and the examples in this document reflect the inconsistency of humans. However, the scoring ignores these premodifiers.

3.1 Expressing Alternative Attribute Values

A vertical bar is being used to separate alternative TYPE attribute values in the answer key. Alternative values will be given when the annotator does not have enough information to make a unique categorization, even considering the context and the annotator's knowledge of the world. This feature will be used sparingly by the annotators.

3.2 Expressing Alternative String Boundaries (ALT Attribute)

The ALT attribute will be used when the tagged string contains one or more strings that should be considered correct for the purposes of scoring the system response. In the following example the annotators have marked the broadest string but allowed the minimum string to be tagged as an alternative.

"all of 1987"

```
<B_TIMEX TYPE="DATE" ALT="1987">all of 1987<E_TIMEX>
```

3.3 Expressing Optional Markup (STATUS Attribute)

When it is not certain that a string should be marked up, the annotator will include the STATUS attribute in the markup to indicate that the markup is optional. The only value of the STATUS attribute is “OPT.”

4 GUIDELINES FOR MARKUP OF EXCEPTIONAL CONSTRUCTIONS

4.1 Conjunction and Elision in Multi-modifier Expressions and Multi-Name/Number Expressions

Conjoined named entities in general are marked separately except for multi-modifier expressions. All cases of multi-modifier expressions are tagged as *single* expressions. However, multi-name, multi-number, and numeric range expressions are marked separately even if there is elision of heads in the expressions.

4.1.1 Multi-modifier Expressions

A single-name expression containing conjoined modifiers with no elision should be marked up as a single expression.

“U.S. Fish and Wildlife Service” (which does NOT mean two entities, i.e. “the U.S. Fish Service” and “the U.S. Wildlife Service”)

```
<B_ENAMEX TYPE="ORGANIZATION">U.S. Fish and Wildlife Service<E_ENAMEX>
```

4.1.2 Multi-name (or Multi-number) Expressions

A conjoined multi-name expression, in which there is elision of the head of one conjunct, should be marked up as separate expressions.

“North and South America”

```
<B_ENAMEX TYPE="LOCATION">North<E_ENAMEX> and <B_ENAMEX
```

```
   TYPE="LOCATION">South America<E_ENAMEX>
```

"bill and susan jones"

```
<B_ENAMEX TYPE="PERSON">bill<E_ENAMEX> and <B_ENAMEX TYPE="PERSON">susan
   jones<E_ENAMEX>
```

A similar case occurs with elision in multi-number expressions:

“10- and 20-dollar bills” (i.e. 10-dollar bills and 20-dollar bills)

```
<B_NUMEX TYPE="MONEY">10<E_NUMEX>- and <B_NUMEX TYPE="MONEY">20-
   dollar<E_NUMEX> bills
```

4.2 Numeric Range Expressions

The subparts of time, date, and numeric range expressions should be marked up separately, even if a portion of a subpart is elided. This applies to both TIMEX and NUMEX expressions.

“175 to 180 million Canadian dollars”

```
<B_NUMEX TYPE="MONEY">175<E_NUMEX> to <B_NUMEX TYPE="MONEY">180 million
Canadian dollars<E_NUMEX>
```

"twelve twenty to three _p_m"

```
<B_TIMEX TYPE="TIME">twelve twenty<E_TIMEX> to <B_TIMEX TYPE="TIME">three
   _p_m<E_TIMEX>
```

“from 1990 through 1992”

```
from <B_TIMEX TYPE="DATE">1990<E_TIMEX> through <B_TIMEX
   TYPE="DATE">1992<E_TIMEX>
```

“from five years to 15 years”

```
from <B_TIMEX TYPE="DURATION">five years<E_TIMEX> to <B_TIMEX
   TYPE="DURATION">15 years<E_TIMEX>
```

“between ten and fifteen percent”

```
between <B_NUMEX TYPE="PERCENT">ten<E_NUMEX> and <B_NUMEX
   TYPE="PERCENT">fifteen percent<E_NUMEX>
```

"from zero to sixty"

```
from <B_NUMEX TYPE="MEASURE">zero<E_NUMEX> to <B_NUMEX
   TYPE="MEASURE">sixty<E_NUMEX>
```

[the context being something like “the car goes from zero to sixty miles per hour”]

4.3 Effects of Tokenization Conventions

The systems must incorporate certain tokenization conventions. These conventions are contained in Appendix A entitled “Tokenization Rules.” The tokenization conventions for newswire text have an impact on the boundaries of the strings to be tagged. For example, the conventions call for treating possessive forms, e.g., “California’s” as multiple tokens, unless there is a name such as “McDonald’s [burger company]” that is inherently possessive.

However, in speech transcriptions, punctuation and capitalization are often lacking and, to make matters worse, interjections and disfluencies occur. The Tokenization Rules will provide examples of markup for all of these transcription cases as well as the newswire cases so that the boundaries of the strings to be tagged are clear. Entities should be tagged even when they include interjections or disfluencies.

In speech transcriptions, some SGML annotations and shortrefs are inserted during the transcription process. These annotations should be included within the named entity markup wherever they occur within a named entity or there is no white space to separate them from the beginning or end of the name; otherwise, they should be left outside of the named entity markup. Examples of markup for all types of these transcription annotations are provided in

Appendix B.

“richard allen %uh men's society”

```
<B_ENAMEX TYPE="ORGANIZATION">richard allen %uh men's society<E_ENAMEX>
```

"wall {laugh street journal"

```
<B_ENAMEX TYPE="ORGANIZATION">wall {laugh street journal<E_ENAMEX>
```

"gloria
<comment>
WMF: should be {ms. | miss}
</comment>
allred"

```
<B_ENAMEX TYPE="PERSON">gloria
<comment>
WMF: should be {ms. | miss}
</comment>
allred<E_ENAMEX>
```

In various text sources there are some special characters used that end up being within the marked string because they are contiguous, but a reader will ignore them. For example, in the Wall Street Journal an @ appears at the beginning of some lines in the headline. In the New York Times News Service articles there are some codes such as “&MD;” which appear and are not always separated by white space from their environment. These will generally be marked up and the scorer will not be able to delete them because of the segmentation problem. Although infrequent, the rule we follow will be to include them if they are string-internal and to exclude them otherwise. It is unlikely that scores will be seriously affected so the scorer will not specially treat these codes.

4.4 Nested Expressions

No nested expressions will be marked. Even in cases where LOCATION (ENAMEX) expressions occur within ENAMEX, TIMEX, and NUMEX expressions, they are not to be tagged. Also, entity names or numbers that appear within ENAMEX tags are *not* to be tagged.

“8:24 a.m. Chicago time”

```
<B_TIMEX TYPE="TIME">8:24 a.m. Chicago time<E_TIMEX>
```

“U.S. $10 million”

```
<B_NUMEX TYPE="MONEY">U.S. $10 million<E_NUMEX>
```

“the U.S. Customs Service”

```
the <B_ENAMEX TYPE="ORGANIZATION">U.S. Customs Service<E_ENAMEX>
```

“4766 Broadway”

```
<B_ENAMEX TYPE="LOCATION">4766 Broadway<E_ENAMEX>
```

5 ENAMEX: SPECIFIC GUIDELINES

5.1 Guidelines That Pertain to All TYPEs (PERSON, LOCATION, ORGANIZATION)

5.1.1 Entity Expressions that Modify Non-taggables

Proper names used as modifiers in complex NPs are to be tagged when it is clear to the annotator from context or the annotator’s world knowledge that the modifier name is that of an organization, person, or location.

"the Clinton government"

```
the <B_ENAMEX TYPE="PERSON">Clinton<E_ENAMEX> government
```

"Treasury bonds and securities"

```
<B_ENAMEX TYPE="ORGANIZATION">Treasury<E_ENAMEX> bonds and securities
```

"U.S. exporters"

```
<B_ENAMEX TYPE="LOCATION">U.S.<E_ENAMEX> exporters
```

"Apple computers"

```
<B_ENAMEX TYPE="ORGANIZATION">Apple<E_ENAMEX> computers
```

"the oklahoma bombing"

```
the <B_ENAMEX TYPE="LOCATION">oklahoma<E_ENAMEX> bombing
```

"a delta jetliner"

```
a <B_ENAMEX TYPE="ORGANIZATION">delta<E_ENAMEX> jetliner
```

"Chrysler division"

```
<B_ENAMEX TYPE="ORGANIZATION">Chrysler<E_ENAMEX> division
```

When capitalization information is available, note that uncapitalized, organization-designating common or proper nouns such as “division” in the phrase “Chrysler division” are *not* considered part of an entity name. Whenever capitalization information is not available or is unreliable, as in the case of speech transcripts (see Appendix A), then organization-designating common or proper nouns *are* considered part of the name.

"chrysler division"

```
<B_ENAMEX TYPE="ORGANIZATION">chrysler division<E_ENAMEX>
```

"the gallup organization"

```
the <B_ENAMEX TYPE="ORGANIZATION">gallup organization<E_ENAMEX>
```

"machinists union"

```
<B_ENAMEX TYPE="ORGANIZATION">machinists union<E_ENAMEX>
```

5.1.2 Entity-Strings Embedded in Entity Expressions

In some cases, taggable multi-word strings will contain entity name substrings; such multi-word strings are not decomposable; therefore, the substrings are not to be tagged.

“Arthur Anderson Consulting”

```
<B_ENAMEX TYPE="ORGANIZATION">Arthur Anderson Consulting<E_ENAMEX>
```

[no markup for "Arthur Anderson" alone]

“Boston Chicken Corp.”

```
<B_ENAMEX TYPE="ORGANIZATION">Boston Chicken Corp.<E_ENAMEX>
```

[no markup for "Boston" alone]

“U.S. Fish and Wildlife Service”

```
<B_ENAMEX TYPE="ORGANIZATION">U.S. Fish and Wildlife Service<E_ENAMEX>
```

[no markup for "U.S." alone]

"pennsylvania state nurses association"

```
<B_ENAMEX TYPE="ORGANIZATION">pennsylvania state nurses
   association<E_ENAMEX>
```

[no markup for "pennsylvania" alone]

5.1.3 Entity Expressions that Possess Other Entity Expressions

In a possessive construction, the possessor and possessed ENAMEX substrings should be tagged separately.

“Temple University's Graduate School of Business”

```
<B_ENAMEX TYPE="ORGANIZATION">Temple University<E_ENAMEX>'s <B_ENAMEX
   TYPE="ORGANIZATION">Graduate School of Business<E_ENAMEX>
```

“California's Silicon Valley”

```
<B_ENAMEX TYPE="LOCATION">California<E_ENAMEX>'s <B_ENAMEX
   TYPE="LOCATION">Silicon Valley<E_ENAMEX>
```

"_g_m’s hughes electronics"

```
<B_ENAMEX TYPE="ORGANIZATION">_g_m<E_ENAMEX>’s <B_ENAMEX
   TYPE="ORGANIZATION">hughes electronics<E_ENAMEX>
```

“Canada’s Parliament”

```
<B_ENAMEX TYPE="LOCATION">Canada<E_ENAMEX>'s <B_ENAMEX
   TYPE="ORGANIZATION">Parliament<E_ENAMEX>
```

5.1.4 Entity Expression Aliases

5.1.4.1 Taggable Aliases

Generally, aliases for entities are to be tagged. Taggable aliases will include the following forms of entity names:

- Acronyms, formed from the initial letter(s) or syllable(s) of successive or major parts of a compound term. Note that speech examples of acronyms may appear in a non-standard format. For example:

  "IBM" [alias for International Business Machines Corp.]

  ```
  <B_ENAMEX TYPE="ORGANIZATION">IBM<E_ENAMEX>
  ```

  "PACTEL" [alias for Pacific Telesys, i.e. Pacific Telephone Systems]

  ```
  <B_ENAMEX TYPE="ORGANIZATION">PACTEL<E_ENAMEX>
  ```

  "_a_t and _t"

  ```
  <B_ENAMEX TYPE="ORGANIZATION">_a_t and _t<E_ENAMEX>
  ```

  "_n double _a_c_p urban league"

  ```
  <B_ENAMEX TYPE="ORGANIZATION">_n double _a_c_p<E_ENAMEX> <B_ENAMEX
     TYPE="ORGANIZATION">urban league<E_ENAMEX>
  ```

- Nicknames and other aliases are tagged when they are established alternate ways of referring to an entity; if the annotator does not recognize the status of the nickname, it may be possible to determine from context whether the nickname is “established” or not. Nicknames and other neologisms that are not commonly used to refer to an entity are not to be tagged (see section **5.2.7 Miscellaneous Personal Non-taggables**). Taggable examples include:

  "Big Blue" [alias for International Business Machines Corp.]

  ```
  <B_ENAMEX TYPE="ORGANIZATION">Big Blue<E_ENAMEX>
  ```

  "Big Board" [alias for New York Stock Exchange]

  ```
  <B_ENAMEX TYPE="ORGANIZATION">Big Board<E_ENAMEX>
  ```

  "Mr. Fix-It" [nickname for candidate for head of the CIA]

  ```
  Mr. <B_ENAMEX TYPE="PERSON">Fix-It<E_ENAMEX>
  ```

"the Big Apple" [nickname for New York City]

```
<B_ENAMEX TYPE="LOCATION">the Big Apple<E_ENAMEX>
```

"the garden state" [nickname for New Jersey]

```
<B_ENAMEX TYPE="LOCATION">the garden state<E_ENAMEX>
```

"the big easy" [nickname for New Orleans]

```
<B_ENAMEX TYPE="LOCATION">the big easy<E_ENAMEX>
```

- Truncated Names, provided that the resulting form is clearly a proper noun referring to a specific entity, for example in:

"Red Sox" [alias for the Boston Red Sox]

```
<B_ENAMEX TYPE="ORGANIZATION">Red Sox<E_ENAMEX>
```

"Sears" [alias for Sears Roebuck and Co.]

```
<B_ENAMEX TYPE="ORGANIZATION">Sears<E_ENAMEX>
```

- Certain metonyms, herein designated “proper” metonyms, which chiefly include references to an organization based on the name of a unique structure or facility in which the organization holds office. The association between the name and the organization should be idiosyncratic enough to justify its inclusion in a dictionary definition of the term (in contrast with "common" metonyms, discussed below), as a kind of nickname for the organization. Some examples follow.

"The White House announced ..." [alias for the U.S.president's executive organization]

```
The <B_ENAMEX TYPE="ORGANIZATION">White House<E_ENAMEX> announced ...
```

"The Pentagon announced..."

```
The <B_ENAMEX TYPE="ORGANIZATION">Pentagon<E_ENAMEX> announced ...
```

- Metonyms, herein designated “common” metonyms, that reference political, military, athletic, and other organizations by the name of a city, country, or other associated location. In these cases, the association between the name’s semantic type and the organization is sufficiently predictable and non-idiosyncratic as to preclude a dictionary gloss; hence the name should be tagged as a LOCATION, not as an ORGANIZATION. Some examples of “common” metonyms follow.

"Germany invaded Poland in 1939."

```
<B_ENAMEX TYPE="LOCATION">GERMANY<E_ENAMEX> invaded <B_ENAMEX
   TYPE="LOCATION">Poland<E_ENAMEX> in …
```

"Baltimore defeated the Yankees by a score of 4 to 3.

```
<B_ENAMEX TYPE="LOCATION">Baltimore</LOCATION> defeated the <B_ENAMEX
   TYPE="ORGANIZATION">Yankees</ORGANIZATION> ...
```

Note that links from LOCATION-tagged names to organizations (e.g. “Baltimore” to the “Baltimore Orioles” baseball team) are left to occur, along with anaphora-resolution, at a processing level higher than Named Entity tagging.

5.1.4.2 Non-taggable Aliases

The following forms of entity names will NOT be tagged:

- Common nouns, including pronouns, used in anaphoric reference to taggable entity names, such as

"IBM announced that the company would lay off ..."
[no markup for "the company"]

- Aliases that refer to broad industrial sectors, political power centers, etc., rather than to specific organizations. For example, do not tag “Wall Street” as an alias for the U.S. stock market, “Japan Incorporated” as an alias for

Japanese Industries, “Uncle Sam” and “Washington” as aliases for the U.S. government, or “Capitol Hill” as an alias for the Congress, since these do not refer to specific organizations. The “Ivy League” refers to a specific set of universities, but does not seem to be a specific organization in its own right. Similarly, the “Axis” (WWII Germany-Japan-Italy) and the “Iron Curtain countries” are aliases for finite sets of entities, but not for specific organizations with corporation-like infrastructures. Note also that “Wall Street” and “Capitol Hill” will not be marked as location, since the terms as they are **commonly** used no longer refer to the actual geographical locations. However, “Washington” will be tagged as location, and “Uncle Sam” will be tagged as a fictional person.

5.1.5 Quotation Marks Around an Entity Name

Quotes are included in the tag if they appear within an entity's name,

“Vito “The Godfather” Corleone”

```
<B_ENAMEX TYPE="PERSON">Vito "The Godfather" Corleone<E_ENAMEX>
```

but not if they bound the name:

“Corleone, also known as “The Godfather,” was the victim of a mob “hit”...”

```
...also known as "<B_ENAMEX TYPE="PERSON">The Godfather<E_ENAMEX>,” was
   the...
```

Such quotation marks are conventional in written language, but in spoken language a paraphrase is more likely to appear. For example,

"vito corleone, known as the godfather"

```
<B_ENAMEX TYPE="PERSON">vito corleone<E_ENAMEX>, known as <B_ENAMEX
   TYPE="PERSON">the godfather<E_ENAMEX>
```

However, the following phrases should be tagged as single expressions.

"vito the godfather"

```
<B_ENAMEX TYPE="PERSON">vito the godfather<E_ENAMEX>
```

"vito the godfather corleone"

```
<B_ENAMEX TYPE="PERSON">vito the godfather corleone<E_ENAMEX>
```

5.1.6 The Definite Article in an Entity Name

When a definite article is commonly associated with an entity name, it also must be tagged.

“when The Godfather ordered the hit”

```
when <B_ENAMEX TYPE="PERSON">The Godfather<E_ENAMEX> ordered the hit
```

"The Hague"

```
<B_ENAMEX TYPE="LOCATION">The Hague<E_ENAMEX>
```

However, the scoring program ignores a certain list of premodifiers as specified (see section 3) which may make the scoring in some of these cases more lenient than this rule implies.

5.1.7 Non-decomposable Names

Complex proper names that are not to be marked (because the whole name does not refer to a currently recognized ENAMEX entity) cannot be decomposed. Entity names used as modifiers within these complex proper names are *not* to be marked separately. Non-taggables named after persons should not have the person’s name marked. A company name which is part of a facility name should not be tagged separately. (However, a company name within a product name can be extracted. See section **5.4.4 Decomposable Product Names**). Note also that proper metonyms are a separate case (see section 5.1.4.1).

"China Film Festival"
[no markup, this is the name of an event]

"Qualcomm Stadium"
[no markup, this name refers to the stadium, not the company "Qualcomm"]

"the Nobel Prize"
[no markup, this name refers to the prize itself, not the actual man it was named after]

“Washington Journal”
[Title of a TV show. No markup for location “Washington”]

"the movie ‘Shakespeare’s Sister’"
[no markup of the person “Shakespeare”]

"Chicago Hope"
[no markup of the location “Chicago”]

5.1.8 Miscellaneous Non-taggables

5.1.8.1 Figures of Speech

Figures of speech include expressions such as metaphors or similes or devices such as personification or hyperbole. Such expressions which use an otherwise taggable ENAMEX expression are not to be tagged.

"the mark fuhrman of corporate america"
[no markup; the name as used here does not refer to the individual with this name, but to some behavior associated with him.]

"(this candidate) will pull a bush"
[no markup, same comment as above.]

"only god knows what will happen"
[no markup. See section 5.2.4 for comments on religious figures.]

Other figures of speech include plural names that do not identify a single, unique entity.

"the campbell soups of the world"
[no markup]

5.1.8.2 Non-taggable Proper Names

Miscellaneous types of proper names that are *not* to be tagged as ENAMEX include names of events, products, media (such as TV and radio shows, movies, and books), and treaties. (For information on the treatment of facilities, see section 5.4.2.5 below.)

"are you going to the olympics"
[no markup]

"tom selleck’s shirt from magnum _p_i"

```
<B_ENAMEX TYPE="PERSON">tom selleck<E_ENAMEX>’s shirt from magnum _p_i
```

[no markup for magnum _p_i]

"this morning on marketplace"
[no markup]

"the movie independence day"
[no markup]

"the _c_t_b_t may be signed in the future”
[no markup]

5.2 Guidelines That Pertain Only to PERSON

This entity type includes not only humans both “real” and “fictional,” but also other fictional or real non-human individuals. The reasoning for including these non-humans in the PERSON type is given in the following sections. We choose not to divide this type into different sub-types, leaving such resolution and sense-evaluation to higher-level processing.

5.2.1 Titles of PERSONs

5.2.1.1 Titles vs. Generational Designators

Titles such as “Mr.” and role names such as “President” are *not* considered part of a person name. However, appositives such as “Jr.,” “Sr.” and “III” *are* considered part of a person name. These will be spelled out rather than abbreviated in speech transcriptions.

"Mr. Harry Schearer"
```
Mr. <B_ENAMEX TYPE="PERSON">Harry Schearer<E_ENAMEX>
```

"Secretary Robert Mosbacher"
```
Secretary <B_ENAMEX TYPE="PERSON">Robert Mosbacher<E_ENAMEX>
```

"John Doe, Jr."
```
<B_ENAMEX TYPE="PERSON">John Doe, Jr.<E_ENAMEX>
```

"mister bettelheim"
```
mister <B_ENAMEX TYPE="PERSON">bettelheim<E_ENAMEX>
```

"the reverend %uh jackson mentioned this"
```
the reverend %uh <B_ENAMEX TYPE="PERSON">jackson<E_ENAMEX> mentioned this
```

5.2.1.2 Entities that Modify Persons/Titles

Entity names modifying a person or their title/role are to be tagged.

"Mips Vice President John Hime" [Mips is the name of a computer company]
```
<B_ENAMEX TYPE="ORGANIZATION">Mips<E_ENAMEX> Vice President <B_ENAMEX
   TYPE="PERSON">John Hime<E_ENAMEX>
```

"Treasury Secretary"
```
<B_ENAMEX TYPE="ORGANIZATION">Treasury<E_ENAMEX> Secretary
```

"the U.S. Vice President"
```
the <B_ENAMEX TYPE="LOCATION">U.S.<E_ENAMEX> Vice President
```

5.2.2 Family Entity Expressions

Family names are to be tagged as PERSON.

"the Kennedy family"

```
the <B_ENAMEX TYPE="PERSON">Kennedy<E_ENAMEX> family
```

"the Kennedys"

```
the <B_ENAMEX TYPE="PERSON">Kennedys<E_ENAMEX>
```

[alternate form of identification of the Kennedy family entity]

5.2.3 Names of Animals

Animal names are to be tagged as PERSON. References to animals and other non-human characters can occur frequently in some types of speech transcriptions. These are unique figures, sometimes representative of some other entity as a type of spokesperson.

"Buddy, the current president’s dog, went to the vet today."

```
<B_ENAMEX TYPE="PERSON">Buddy<E_ENAMEX>, the current president’s dog, went
   to the vet today.
```

5.2.4 Saints and other Religious Figures

Although religious titles or specifiers such as “saint,” “prophet,” “imam,” or “archangel” are not be tagged, the proper name will be tagged as a PERSON. This practice becomes more significant in marking up speech transcriptions, due to peculiarities of speech habits or patterns.

"St. Christopher is the patron of"

```
St. <B_ENAMEX TYPE="PERSON">Christopher<E_ENAMEX> is the patron of
```

"the imam reza"

```
the imam <B_ENAMEX TYPE="PERSON">reza<E_ENAMEX>
```

References to “God” will be taken to be the “name” of this entity for tagging purposes. If it is used as a descriptor, rather than a name, it will not be tagged. Note that capitalization information may not be available in speech transcripts.

"if you believe in god you must..."

```
if you believe in <B_ENAMEX TYPE="PERSON">god<E_ENAMEX> you must...
```

"although he felt like he was a god he..."

[no markup]

5.2.5 Fictional Characters

Names of fictional characters are to be tagged; however, character names used as TV show titles will not be tagged when they refer to the show, rather than the character name. This is in keeping with the rule which excludes markup of person names in artifacts that were named after a person (see section 1.5.2.7).

"batman has become a popular icon"

```
<B_ENAMEX TYPE="PERSON">batman<E_ENAMEX> has become a popular icon
```

"adam west’s costume from batman the _t_v series"

```
<B_ENAMEX TYPE="PERSON">adam west<E_ENAMEX>’s costume from batman the _t_v
   series
```

5.2.6 Fictional Animals and Non-human Characters

Fictional animals are a specific type of fictional character, and as such should be tagged. References to animals and other non-human characters can occur frequently in some types of speech transcriptions. These are unique figures, sometimes representative of an organization as a type of spokesperson.

"morris the cat"

```
<B_ENAMEX TYPE="PERSON">morris the cat<E_ENAMEX>
```

"that famous advertising icon speedy"

```
that famous advertising icon <B_ENAMEX TYPE="PERSON">speedy<E_ENAMEX>
```

"snuggle the fabric softener bear"

```
<B_ENAMEX TYPE="PERSON">snuggle<E_ENAMEX> the fabric softener bear
```

5.2.7 Miscellaneous Personal Non-taggables

Miscellaneous types of proper names that are not to be tagged as PERSON include: individuals identified by their political affiliation, laws named after people, court cases named after people, weather formations, and diseases/ prizes named after people.

"The Republican stepped into the voting booth."
[no markup]

"the Gramm-Rudman amendment"
[no markup]

"Alzheimer's"
[no markup]

"the Nobel Prize"
[no markup]

"tropical storm arthur"
[no markup]

"in the case of joe castano versus the tobacco growers of america"
[no markup]

However, the following example refers to the name of the person involved in the lawsuit, not the name of the lawsuit itself:

"in the castano suit, attorneys argued"

```
in the <B_ENAMEX TYPE="PERSON">castano<E_ENAMEX> suit, attorneys argued
```

Nicknames and other neologisms that are *not* commonly used to refer to an entity are not to be tagged. Nicknames and other aliases are tagged only when they are established ways of referring to an entity (see section **5.1.4 Entity Expression Aliases**) .

"and al gore will legally change his first name to planet "

```
and <B_ENAMEX TYPE="PERSON">al gore<E_ENAMEX> will legally change his first
   name to planet
```

[no markup for planet]

5.3 Guidelines That Pertain Only to LOCATION

The TYPE LOCATION applies to entities representing either geographical, political, or astronomical locations. Examples of strings that are tagged as LOCATION include: named heavenly bodies, continents, countries, provinces, counties, cities, regions, districts, towns, villages, neighborhoods, airports, military bases, railways, railroads, highways, bridges, street names, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, fictional or mythical locations, and certain structures, such as the Eiffel

Tower and Washington Monument, that were built primarily as monuments.

"it will be called the jim and susan mcdougal bridge to honor their contributions"

```
"it will be called the <B_ENAMEX TYPE="LOCATION">jim and susan mcdougal
   bridge<E_ENAMEX> to honor their contributions"
```

"violence in the housing projects florida and desire"

```
violence in the housing projects <B_ENAMEX
   TYPE="LOCATION">florida<E_ENAMEX> and <B_ENAMEX
   TYPE="LOCATION">desire<E_ENAMEX>
```

"fremont canyon and laguna peak"

```
<B_ENAMEX TYPE="LOCATION">fremont canyon<E_ENAMEX> and <B_ENAMEX
   TYPE="LOCATION">laguna peak<E_ENAMEX>
```

"the silk road railway"

```
the <B_ENAMEX TYPE="LOCATION">silk road railway<E_ENAMEX>
```

"independence mall"

```
<B_ENAMEX TYPE="LOCATION">independence mall<E_ENAMEX>
```

"created a backup at O'Hare International Airport"

```
created a backup at <B_ENAMEX TYPE="LOCATION">O'Hare International
   Airport<E_ENAMEX>
```

Even if the name of the airport refers to the organization or business of the airport and not its location or facilities, it is still marked as a LOCATION:

"Massport, which owns and operates Logan, defended the attempts..."

```
<B_ENAMEX TYPE="ORGANIZATION">Massport<E_ENAMEX>, which owns and operates
   <B_ENAMEX TYPE="LOCATION">Logan<E_ENAMEX>, defended the attempts...
```

5.3.1 Possible Embedded Locative Entity-Strings

A location name following an organization name may or may not be part of the organization name proper. The annotation in the answer key will follow these guidelines: (1) If the location name is preceded by "in" or "at," the location name will be tagged separately from the organization name, unless world knowledge dictates otherwise (see note below). If the location name is preceded by "of" or no preposition at all, then determine tagging as follows: (2) If there is an organization designator or organization-designating common or proper noun, it marks the end of the organization name; (3) if there is no organization designator, or if the organization is a bank or university, the location name is part of the organization name; (4) if the organization is a radio station for which the location name is the location of the radio station's transmitter, the location will be tagged separately.

"Hyundai of Korea, Inc."

```
<B_ENAMEX TYPE="ORGANIZATION">Hyundai of Korea, Inc.<E_ENAMEX>
```

"Hyundai, Inc. of Korea"

```
<B_ENAMEX TYPE="ORGANIZATION">Hyundai, Inc.<E_ENAMEX> of <B_ENAMEX
   TYPE="LOCATION">Korea<E_ENAMEX>
```

"the distilled spirits council of the united states"

```
the <B_ENAMEX TYPE="ORGANIZATION">distilled spirits council<E_ENAMEX> of
   the <B_ENAMEX TYPE="LOCATION">united states<E_ENAMEX>
```

"McDonald's of Japan"

```
<B_ENAMEX TYPE="ORGANIZATION">McDonald's of Japan<E_ENAMEX>
```

"faculty of the university of san diego"

```
faculty of the <B_ENAMEX TYPE="ORGANIZATION">university of san diego<E_ENAMEX>
```

"Lockheed Martin, North Asia"

```
<B_ENAMEX TYPE="ORGANIZATION">Lockheed Martin, North Asia<E_ENAMEX>
```

"_w_g_b_h boston"

```
<B_ENAMEX TYPE="ORGANIZATION">_w_g_b_h<E_ENAMEX> <B_ENAMEX
   TYPE="LOCATION">boston<E_ENAMEX>
```

Note: locatives are often intrinsic to names of universities and should be included in the organization name when world knowledge dictates, such as in the following:

"american university in cairo"

```
<B_ENAMEX TYPE="ORGANIZATION">american university in cairo<E_ENAMEX>
```

“World knowledge” includes the application of using a commonly-used acronym to help determine extent of the name. For instance:

“University of California, Los Angeles”

```
<B_ENAMEX TYPE="ORGANIZATION">University of California, Los
   Angeles<E_ENAMEX>
```

[Since the university is commonly referred to as “UCLA,” the proper ENAMEX expression must use the entire extent indicated by that acronym as the entity name.]

5.3.2 Locative Entity Expressions Tagged in Succession

Compound expressions in which place names are listed in succession, with or without a separating comma, are to be tagged as separate instances of LOCATION.

“Kaohsiung, Taiwan”

```
<B_ENAMEX TYPE="LOCATION">Kaohsiung<E_ENAMEX>, <B_ENAMEX
   TYPE="LOCATION">Taiwan<E_ENAMEX>
```

“Washington, D.C.”

```
<B_ENAMEX TYPE="LOCATION">Washington<E_ENAMEX>, <B_ENAMEX TYPE="LOCATION">
   D.C.<E_ENAMEX>
```

“newark new jersey”

```
<B_ENAMEX TYPE="LOCATION">newark<E_ENAMEX> <B_ENAMEX TYPE="LOCATION">new
   jersey<E_ENAMEX>
```

“kokomo indiana”

```
<B_ENAMEX TYPE="LOCATION">kokomo<E_ENAMEX> <B_ENAMEX
   TYPE="LOCATION">indiana<E_ENAMEX>
```

5.3.3 Locative Designators and Specifiers

Designators that are integrally associated with a place name are to be tagged as part of the name. For example, include in the tagged string the word “River” in the name of a river, “Mountain” in the name of a mountain, “City” in the name of a city, etc., if such words are contained in the string.

"Mississippi River"

```
<B_ENAMEX TYPE="LOCATION">Mississippi River<E_ENAMEX>
(not: <B_ENAMEX TYPE="LOCATION">Mississippi<E_ENAMEX> River)
```

Separability criteria are useful in speech transcriptions where capitalization is unavailable. Note that “mississippi river valley” means the same thing as “the valley of the mississippi river,” while “the ohio valley” is not the same as “the valley of ohio.” If the extent of the location name cannot be determined from world knowledge or from the appropriate reference atlas, then the common-noun designator will be included in the location name.

"ohio valley"

```
<B_ENAMEX TYPE="LOCATION">ohio valley<E_ENAMEX>
```

"_l_a basin"

```
<B_ENAMEX TYPE="LOCATION">_l_a basin<E_ENAMEX>
```

"texas panhandle"

```
<B_ENAMEX TYPE="LOCATION">texas panhandle<E_ENAMEX>
```

5.3.3.1 Locative Non-taggables: The Postposed Partitive Specifier

Do not include in the tagged string common noun phrases functioning as partitive-type locative specifiers directly after LOCATION names, such as:

"Mississippi River west bank" (west bank of the Mississippi River)

```
<B_ENAMEX TYPE="LOCATION">Mississippi River<E_ENAMEX> west bank
```

"mississippi river valley"

```
<B_ENAMEX TYPE="LOCATION">mississippi river<E_ENAMEX> valley
```

5.3.3.2 Exceptional Locative Specifiers Used as Entity Expressions

Note that, due to the political significance of the Jordan River's west bank, the term “West Bank” assumes the status of a named entity expression. A similar example is the term “Left Bank” (of the Seine River) as a name for an area of Paris. Use world knowledge or the appropriate reference atlas to determine whether such a term is being used as a specifying non-taggable following a place name, or as an entity expression (a proper noun) representing a particular LOCATION.

5.3.4 Transnational and Subnational Region Names

5.3.4.1 Transnational Locative Entity Expressions

Tag names of continents (“Africa”), multi-country sub-continental regions (“Eastern Europe,” “Sub-Saharan Africa”), and multi-country trans-continental regions (“the Middle East,” “the Pacific Rim”).

5.3.4.2 Subnational Region Names

Do not tag names of sub-national regions when referenced only by compass-point modifiers. Do not tag “the South” or the “mid-West,” analogies to “the Middle East” notwithstanding, because, unlike the latter term, their referential value varies from country to country. For example,

“the Southwest region”
[no markup]

"the northeast"
[no markup]

Do tag names of sub-national but specific regions if they are identifiable even when the name is disassociated from context. Examples include “the Ruhr,” “the Rockies,” “the Auvergne,” and “Amazonia.” Note that these names generally straddle, or lie within, geo-political jurisdictions such as states or provinces.

5.3.5 Time Modifiers of Locative Entity Expressions

Historic-time modifiers (“former,” “present-day”) are not to be included in tagged expressions. These are used as ad hoc modifiers that are readily separable from the name.

“urgently transported out of the former Soviet republic of Georgia”

```
urgently transported out of the former <B_ENAMEX
   TYPE="LOCATION">Soviet<E_ENAMEX> republic of <B_ENAMEX
   TYPE="LOCATION">Georgia<E_ENAMEX>
```

"former Soviet Union"

```
former <B_ENAMEX TYPE="LOCATION">Soviet Union<E_ENAMEX>
```

"Gaul (present-day France)"

```
<B_ENAMEX TYPE="LOCATION">Gaul<E_ENAMEX> (present-day <B_ENAMEX
   TYPE="LOCATION">France<E_ENAMEX>)
```

Contrast “Premier of the former Soviet Union” and “formerly Premier of the Soviet Union” to see the separability of these modifiers.

5.3.6 Space Modifiers of Locative Entity Expressions

Directional modifiers (“north,” “south,” “east,” “west,” “upper,” “lower,” and combinations thereof) are taggable only when they are intrinsic parts of a location's official name, as in “Upper Volta” or “North Dakota.”

“lower Manhattan”

```
lower <B_ENAMEX TYPE="LOCATION">Manhattan<E_ENAMEX>
```

"atlanta area"

```
<B_ENAMEX TYPE="LOCATION">atlanta<E_ENAMEX> area
```

"great lakes region"

```
<B_ENAMEX TYPE="LOCATION">great lakes<E_ENAMEX> region
```

Contrast “east Baltimore” and “eastern section of Baltimore”; and “Upper Volta” and “upper section of Volta” to see the separability of these modifiers.

Note that in speech transcriptions without capitalization, this separability criterion is quite useful. If it is not possible to determine either through separability, world knowledge, or the appropriate reference atlas, then the directional modifier should not be included.

```
<B_ENAMEX TYPE="LOCATION">north carolina<E_ENAMEX>
southern <B_ENAMEX TYPE="LOCATION">california<E_ENAMEX>
northern <B_ENAMEX TYPE="LOCATION">california<E_ENAMEX>
central <B_ENAMEX TYPE="LOCATION">asia<E_ENAMEX>
```

5.3.7 Miscellaneous Locative Non-taggables

5.3.7.1 Adjectival Forms of Location Names

Adjectival forms of location names are not to be tagged as LOCATIONs.

"American exporters"
[no markup]

"russian air force"

```
russian <B_ENAMEX TYPE="ORGANIZATION">air force<E_ENAMEX>
```

"zairian rebel coalition"

```
<B_ENAMEX TYPE="ORGANIZATION" STATUS="OPT">zairian rebel
   coalition<E_ENAMEX>
```

[unsure if this is a true organization, so mark it optional]

Note that, as in the last example, an adjective *can* be included within another ENAMEX string if it is indeed part of the entity's name.

5.4 Guidelines That Pertain Only to ORGANIZATION

5.4.1 Corporate or Organization Designators

Corporate or organization designators such as "Co." are considered part of an organization name.

"Bridgestone Sports Co."

```
<B_ENAMEX TYPE="ORGANIZATION">Bridgestone Sports Co.<E_ENAMEX>
```

5.4.2 Miscellaneous ORGANIZATION-type Entity Expressions

Proper names that are to be tagged as ORGANIZATION include stock exchanges, multinational organizations, businesses, TV or radio stations, political parties, religions or religious groups, orchestras, bands, or musical groups, unions, non-generic governmental entity names such as "Congress" or "Chamber of Deputies," sports teams and armies (unless designated only by country names, which are tagged as LOCATION), as well as fictional organizations (to ensure consistency with marking other fictional entities).

"NASDAQ"

```
<B_ENAMEX TYPE="ORGANIZATION">NASDAQ<E_ENAMEX>
```

[a stock exchange]

"European Community"

```
<B_ENAMEX TYPE="ORGANIZATION">European Community<E_ENAMEX>
```

"GOP presidential hopeful"

```
<B_ENAMEX TYPE="ORGANIZATION">GOP<E_ENAMEX> presidential hopeful
```

"the british opposition labor party"

```
the british opposition <B_ENAMEX TYPE="ORGANIZATION">labor party<E_ENAMEX>
```

"practitioners of Santeria worship"

```
practitioners of <B_ENAMEX TYPE="ORGANIZATION">Santeria<E_ENAMEX> worship
```

"the best selling band chumbawamba rocked the awards dinner"

```
the best selling band <B_ENAMEX TYPE="ORGANIZATION">chumbawamba<E_ENAMEX>
   rocked the awards dinner
```

"the mayor who built Candlestick Park for the Giants"

```
the mayor who built Candlestick Park for the <B_ENAMEX
   TYPE="ORGANIZATION">Giants<E_ENAMEX>
```

[a sports team]

"In hockey action, Russia defeated France by a score of 7 to 3."

```
...<B_ENAMEX TYPE="LOCATION">Russia<E_ENAMEX> defeated <B_ENAMEX
   TYPE="LOCATION">France<E_ENAMEX> ...
```

"tonight on _t_ v twelve"

```
tonight on <B_ENAMEX TYPE="ORGANIZATION">_t_v twelve<E_ENAMEX>
```

5.4.2.1 Broadcasting Stations

Stations, channels, and call numbers are to be tagged as ORGANIZATION; however, when a location is given which is the location of the radio station transmitter, the location will be tagged separately:

"_k_c_r_w_f_m eighty nine point nine"

```
<B_ENAMEX TYPE="ORGANIZATION">_k_c_r_w_f_m eighty nine point nine<E_ENAMEX>
```

"_w_g_b_h boston"

```
<B_ENAMEX TYPE="ORGANIZATION">_w_g_b_h<E_ENAMEX> <B_ENAMEX
   TYPE="LOCATION">boston<E_ENAMEX>
```

"ninety one _f_m"

```
<B_ENAMEX TYPE="ORGANIZATION">ninety one _f_m<E_ENAMEX>
```

5.4.2.2 Legislative Bodies

Although Congress can act as either an organization or an event it will be tagged as an ORGANIZATION.

"Congress of Deputies"

```
<B_ENAMEX TYPE="ORGANIZATION">Congress of Deputies<E_ENAMEX>
```

"commencement of the ninety first congress"

```
commencement of the <B_ENAMEX TYPE="ORGANIZATION">ninety first
   congress<E_ENAMEX>
```

5.4.2.3 Event Organizers

Although event names are not to be tagged, even if they refer to events that occur on a regular basis and are associated with institutional structures, the institutional structures themselves -- steering committees, etc. -- should be tagged as ORGANIZATION.

"the U.S. Olympic Committee"

```
<B_ENAMEX TYPE="ORGANIZATION">U.S. Olympic Committee<E_ENAMEX>
```

5.4.2.4 Membership Affiliation

Political or religious parties are sometimes referred to as collections of their members, e.g., “the Moonies,” “Democrats.” Such designators are to be tagged as ORGANIZATION only when they clearly act as such. If the term is being used to refer to an unspecified collection of members of an organization, it should not be tagged. (Note: A human annotator can sometimes determine the meaning by mentally inserting the appropriate head noun to see if the meaning is changed in the given context (e.g., “Republicans” vs. “Republican Party”). When context or world knowledge is not sufficient to determine whether a group is acting as an organization or a person, the name will be tagged as an optional organization.

"The Republicans believe, according to their platform, ..."

```
The <B_ENAMEX TYPE="ORGANIZATION">Republicans<E_ENAMEX> believe, according
   to their platform, ...
```

[Here "Republicans" refers to the "Republican Party,” and should be marked as ORG]

"the iran contra hearings"

```
the <B_ENAMEX TYPE="LOCATION">iran<E_ENAMEX> <B_ENAMEX
   TYPE="ORGANIZATION">contra<E_ENAMEX> hearings
```

"The Republicans were in agreement"

```
The <B_ENAMEX TYPE="ORGANIZATION" STATUS="OPT">Republicans<E_ENAMEX> were
   in agreement
```

[Not clear if this refers to all Republicans as a group, or just a few.]

"The Republicans went to lunch"

[No markup. Clearly this is not a reference to the entire "Republican Party"] "the sandinistas have been mentioned in the press recently"

```
the <B_ENAMEX TYPE="ORGANIZATION" STATUS="OPT">sandinistas<E_ENAMEX> have
been mentioned in the press recently
```

[Not clear if this refers to all Sandinistas as a group, or just a few]

"the contra then turned the gun on himself"

[No markup]

5.4.2.5 ORGANIZATION-related Facilities

Proper names referring to meeting places or places where organizational activities occur (e.g., churches, embassies, factories, hospitals, hotels, museums, universities) will be tagged as ORGANIZATION.

"Finger Lakes Area Hospital Corp."

```
<B_ENAMEX TYPE="ORGANIZATION">Finger Lakes Area Hospital Corp.<E_ENAMEX>
```

"Four Seasons Hotels"

```
<B_ENAMEX TYPE="ORGANIZATION">Four Seasons Hotels<E_ENAMEX>
```

"Unification Church"

```
<B_ENAMEX TYPE="ORGANIZATION">Unification Church<E_ENAMEX>
```

"the White House"

```
the <B_ENAMEX TYPE="ORGANIZATION">White House<E_ENAMEX>
```

"the kremlin"

```
the <B_ENAMEX TYPE="ORGANIZATION">kremlin<E_ENAMEX>
```

"Trinity Lutheran Church"

```
<B_ENAMEX TYPE="ORGANIZATION">Trinity Lutheran Church<E_ENAMEX>
```

"General Hospital"

```
<B_ENAMEX TYPE="ORGANIZATION">General Hospital<E_ENAMEX>
```

"The Empire State Building"

[no markup -- this is a structure that houses many organizations]

"Qualcomm Stadium"

[no markup -- this is a structure that hosts many organizations]

5.4.3 Decomposable Product Names

In cases where the manufacturer and the product are named, the manufacturer will be tagged. The product will not be tagged. Products must be defined loosely to include manufactured products (e.g., vehicles), as well as computed products (e.g., stock indexes) and media products (e.g., television shows).

"Ford Taurus"

```
<B_ENAMEX TYPE="ORGANIZATION">Ford<E_ENAMEX> Taurus
```

"dow jones industrial average"

```
<B_ENAMEX TYPE="ORGANIZATION">dow jones<E_ENAMEX> industrial average
```

"in this _n_b_c poll"

```
in this <B_ENAMEX TYPE="ORGANIZATION">_n_b_c<E_ENAMEX> poll
```

Note that not all product names are decomposable. Only the manufacturer may be extracted from a product name. Other named entities may not. For example:

> "she wanted a 'Hollywood Hair Barbie'"
> [Neither "Hollywood" nor "Barbie" is tagged]

5.4.4 Metonymy

When a publication, regardless of subject matter, "reports," "states," "claims," etc., it is acting as a news source (reporting information). Tag news sources (newspapers, radio and TV stations, and news journals) as ORGANIZATION even when they function as artifacts. This is done because the same name is often used to refer to both the publication and the publisher. To avoid having annotators make the distinction between usage as artifact or organization, both usages are tagged. Moreover, publications often function agentively as ORGANIZATIONs.

"in the wall {laugh street journal today"

```
in the <B_ENAMEX TYPE="ORGANIZATION">wall {laugh street journal<E_ENAMEX>
   today
```

"this morning's new york times"

```
this morning's <B_ENAMEX TYPE="ORGANIZATION">new york times<E_ENAMEX>
```

"tonight on _t_ v twelve"

```
tonight on <B_ENAMEX TYPE="ORGANIZATION">_t_v twelve<E_ENAMEX>
```

Note that TV stations differ from TV shows, the latter not being taggable:

> "more about that tonight on News at Eleven"
> [not tagged]

Similar to publications, metonymy occurs frequently with stock indices and stock markets. Tag these as ORGANIZATION even when they refer to numeric values published by or representing the organization. For example:

"Shares of _J_P Morgan leading the Dow lower"

```
Shares of <B_ENAMEX TYPE="ORGANIZATION">_J_P Morgan<E_ENAMEX> leading the
   <B_ENAMEX TYPE="ORGANIZATION">Dow<E_ENAMEX> lower
```

"Right now the NASDAQ is down thirteen points"

```
Right now the <B_ENAMEX TYPE="ORGANIZATION">NASDAQ<E_ENAMEX> is down
   <B_NUMEX TYPE="MONEY">thirteen points<e_numex>
```

5.4.5 Generic ORGANIZATION-like Non-taggables

Generic entity names such as "the police" and "the government," are not to be tagged. Without capitalization in speech transcriptions, however, it becomes more difficult to distinguish generic non-taggables from true organizations. Non-taggable examples:

> "this country has never been invaded by an army"
> [no markup]

"general of the army"

```
general of the <B_ENAMEX TYPE="ORGANIZATION">army<E_ENAMEX>
```

> "state police do not have jurisdiction in other states"
> [no markup]

While the preceding are descriptive phrases, the following appears to be a named entity, and thus is tagged as

ORGANIZATION:

“louisiana state police have found”

```
<B_ENAMEX TYPE="ORGANIZATION">louisiana state police<E_ENAMEX> have found
```

6 TIMEX: SPECIFIC GUIDELINES

6.1 Introduction

Only “absolute” time expressions are to be tagged in NE99. The TIME type is defined as a temporal unit shorter than a full day, such as second, minute, or hour. The DATE sub-type is a temporal unit of a full day or longer. An additional TIMEX type is DURATION, which captures durations of time.

6.2 Absolute Temporal Expressions - TIME & DATE

To be considered an absolute time expression, the expression must indicate a specific segment of time, as follows:

TIME-tagged expressions

- An expression of minutes must indicate a particular minute and hour, such as “20 minutes after 10” (not “a few minutes after the hour,” “a few minutes after 10,” or “20 minutes after the hour”).
- An expression of hours must indicate a particular hour, such as “midnight,” “twelve o'clock noon,” “noon” (not “mid-day” or “morning” by themselves).

DATE-tagged expressions

- An expression of days must indicate a particular day, such as “Monday,” “10th of October” (not “first day of the month”).
- An expression of seasons must indicate a particular season, such as “autumn” (not “next season”).
- An expression of financial quarters or halves of the year must indicate which quarter or half, such as “fourth quarter,” “first half.” Note that there are no proper names, per se, representing these time periods. Nonetheless, these types of time expressions are important in the business domain and are therefore to be tagged.
- An expression of years must indicate a particular year, such as “1995” (not “the current year”).
- An expression of decades must indicate a particular decade, such as “1980s” (not “the last 10 years”).
- An expression of centuries must indicate a particular century, such as “the 20th century” (not “this century”).

Temporal expressions are to be tagged as a single item. Contiguous subparts (month/day/year) are not to be separately tagged unless they are taggable expressions of two distinct TIMEX sub-types (date followed by time or time followed by date).

“twelve o'clock noon”

```
<B_TIMEX TYPE="TIME">twelve o'clock noon<E_TIMEX>
```

“four o'clock in the morning”

```
<B_TIMEX TYPE="TIME">four o'clock in the morning<E_TIMEX>
```

“5 p.m. EST”

```
<B_TIMEX TYPE="TIME">5 p.m. EST<E_TIMEX>
```

“January 1990”

```
<B_TIMEX TYPE="DATE">January 1990<E_TIMEX>
```

“fiscal 1989”

```
<B_TIMEX TYPE="DATE">fiscal 1989<E_TIMEX>
```

“the autumn report”

```
the <B_TIMEX TYPE="DATE">autumn<E_TIMEX> report
```

“third quarter of 1991”

```
<B_TIMEX TYPE="DATE">third quarter of 1991<E_TIMEX>
```

“the fourth quarter ended Sept. 30”

```
<B_TIMEX TYPE="DATE">the fourth quarter ended Sept. 30<E_TIMEX>
```

“the first half of fiscal 1990”

```
<B_TIMEX TYPE="DATE">the first half of fiscal 1990<E_TIMEX>
```

“first-half profit”

```
<B_TIMEX TYPE="DATE">first-half<E_TIMEX> profit
```

“it’s uh four forty out here”

```
it’s uh <B_TIMEX TYPE="TIME">four forty<E_TIMEX> out here
```

“six oh six”

```
<B_TIMEX TYPE="TIME">six oh six<E_TIMEX>
```

“cars popular in the fifties include”

```
cars popular in the <B_TIMEX TYPE="DATE">fifties<E_TIMEX> include
```

“4th period” [of a year]

```
<B_TIMEX TYPE="DATE">4th period<E_TIMEX>
```

“February 12, 8 A.M.”

```
<B_TIMEX TYPE="DATE">February 12<E_TIMEX>,<B_TIMEX TYPE="TIME">8
   A.M.<E_TIMEX>
```

“by 9 o'clock Monday”

```
by <B_TIMEX TYPE="TIME">9 o'clock<E_TIMEX> <B_TIMEX
   TYPE="DATE">Monday<E_TIMEX>
```

Determiners that introduce the expressions are not to be tagged. Words or phrases modifying the expressions (such as “around” or “about”) also will not be tagged. Only the actual temporal expression itself is to be tagged.

“around the 4th of May”

```
around the <B_TIMEX TYPE="DATE">4th of May<E_TIMEX>
```

“shortly after the 4th of May”

```
shortly after the <B_TIMEX TYPE="DATE">4th of May<E_TIMEX>
```

6.3 Scope of Temporal Expressions

Absolute time expressions combining numerals and time-unit designators (“A.M.,” "P.M.,” "EST,” etc.), or other subparts associated with a single TIMEX sub-type, are to be tagged as a single item. That is, the subparts (such as numbers and time-units) are not to be tagged separately, even in the case of possessive or partitive constructions.

"twelve o'clock noon"

```
<B_TIMEX TYPE="TIME">twelve o'clock noon<E_TIMEX>
```

"four o'clock in the morning"

```
<B_TIMEX TYPE="TIME">four o'clock in the morning<E_TIMEX>
```

"5 p.m. EST"

```
<B_TIMEX TYPE="TIME">5 p.m. EST<E_TIMEX>
```

"the first half of fiscal 1990"

```
the <B_TIMEX TYPE="DATE">first half of fiscal 1990<E_TIMEX>
```

If there has been elision of a portion of the temporal expression (as is common in speech transcripts), the remaining numeric portion will be marked as a TIMEX expression:

"the debate coverage will begin at eight"

```
the debate coverage will begin at <B_TIMEX TYPE="TIME">eight<E_TIMEX>
```

[a time expression is understood, even though “o’clock” is omitted]

6.4 Temporal Expressions Containing Adjacent Absolute and Relative Strings

When a time expression contains both relative and absolute elements, only the absolute expression is to be tagged. The following examples illustrate some of the ways in which elements of relative and absolute time expressions may appear together.

"July last year"

```
<B_TIMEX TYPE="DATE">July<E_TIMEX> last year
```

"the end of 1991"

```
the end of <B_TIMEX TYPE="DATE">1991<E_TIMEX>
```

"late Tuesday"

```
late <B_TIMEX TYPE="DATE">Tuesday<E_TIMEX>
```

"the mid nineteen eighties"

```
the mid <B_TIMEX TYPE="DATE">nineteen eighties<E_TIMEX>
```

“twelve minutes past the hour”

```
<B_TIMEX TYPE="DURATION">twelve minutes<E_TIMEX> past the hour
```

6.5 Holidays

Special days, such as holidays, that are referenced by name, should be tagged.

"because of the observance of All Saints' Day"

```
 because of the observance of <B_TIMEX TYPE="DATE">All Saints' Day<E_TIMEX>
```

"fat tuesday"

```
<B_TIMEX TYPE="DATE">fat tuesday<E_TIMEX>
```

6.6 Locative Entity-Strings Embedded in Temporal Expressions

Rarely, multiword strings that are to be tagged as TIMEX will contain LOCATION (ENAMEX) substrings. Include these words within the scope of the tagged expression, but do not apply an embedded LOCATION tag.

"1:30 p.m. Chicago time"

```
<B_TIMEX TYPE="TIME">1:30 p.m. Chicago time<E_TIMEX>
```

Note that the above locative entity-string (“Chicago”), plus the word “time,” modifies a contiguous TIMEX expression of the “TIME” sub-type.

Sometimes, however, the phrasing is such that the modification and types are non-contiguously arranged as in “Japan time, 19 February, 8:00 A.M.” but marking three items separately does not represent the modification accurately. In such cases, mark the entire phrase as a single temporal expression as shown in the following:

“Japan time, 19 February, 8:00 A.M.”

```
<B_TIMEX TYPE="TIME">Japan time, 19 February, 8:00 A.M.<E_TIMEX>
```

A locative expression should be tagged separately as LOCATION if it is not contiguous to the "TIMEX" type expression, as in:

"In Japan, it would have occurred on 19 February, 8:00 A.M."

```
In <B_ENAMEX TYPE="LOCATION">Japan<E_ENAMEX>, it would have occurred on
   <B_TIMEX TYPE="DATE">19 February<E_TIMEX>, <B_TIMEX TYPE="TIME">8:00
   A.M.<E_TIMEX>
```

6.7 Temporal Expressions Based on Alternate Calendars

Temporal expressions in terms of alternate calendars, such as fiscal years, the Hebrew calendar, Julian dates and "Star Date," will generally be marked up in accordance with the above guidelines for DATE.

6.8 DURATION

TIMEX of type DURATION expressions refer to periods of time. In broadcast news transcripts, they must have the form “[numeral] [time-unit]” where “numeral” is defined as any whole number, fraction, or decimal. Indefinite articles, and the words “few,” “couple,” and “several” are allowable substitutes for numerals in determining whether something counts as a DURATION expression. Numerals, “few,” “couple,” or “several” are included with the time-unit word within the scope of the tag, but indefinite articles are not.

Here is sampling of DURATION expressions:

"it took one night"

```
it took <B_TIMEX TYPE="DURATION">one night<E_TIMEX>
```

“it lasted nearly two hours”

```
it lasted nearly <B_TIMEX TYPE="DURATION">two hours<E_TIMEX>
```

“on and off for twentyfour years”

```
on and off for <B_TIMEX TYPE="DURATION">twentyfour years<E_TIMEX>
```

“within ninety minutes”

```
“within <B_TIMEX TYPE="DURATION">ninety minutes<E_TIMEX>
```

"waited half an hour"

```
waited <B_TIMEX TYPE="DURATION">half an hour<E_TIMEX>
```

“two years ago”

```
<B_TIMEX TYPE="DURATION">two years<E_TIMEX> ago
```

“three to four minutes”

```
<B_TIMEX TYPE="DURATION">three<E_TIMEX> to <B_TIMEX TYPE="DURATION">four
   minutes<E_TIMEX>
```

“couple of minutes”

```
<B_TIMEX TYPE="DURATION">couple of minutes<E_TIMEX>
```

“a few days ago”

```
a <B_TIMEX TYPE="DURATION">few days<E_TIMEX> ago
```

“in one moment”

```
in <B_TIMEX TYPE="DURATION">one moment<E_TIMEX>
```

“it happened in less than a decade”

```
it happened in less than a <B_TIMEX TYPE="DURATION">decade<E_TIMEX>
```

Even pre-nominal modifiers are tagged as TIMEX type DURATION if the expression is of the form “[numeral] [time unit]”. (Non-temporal and non-monetary pre-nominal numeric expressions are tagged as NUMEX type MEASURE -- see section 7.4.3.2 Measurement Phrases).

"one night stand"

```
<B_TIMEX TYPE="DURATION">one night<E_TIMEX> stand
```

In newswire data, the time unit may not be present, but the expression is still markable. For example:

“Prokurorov was 1:18:9 ahead of Hjelmeseth”

```
<B_ENAMEX TYPE="PERSON">Prokurorov<E_ENAMEX> was <B_TIMEX
   TYPE="DURATION">1:18.9<E_TIMEX> ahead of < B_ENAMEX
   TYPE="PERSON">Hjelmeseth< E_ENAMEX >
```

6.8.1   Scope of DURATION Expressions

Dimension words or prepositional phrases following DURATION expressions should *not* be included within the scope of the tags. For example:

“the meeting was two hours long”

```
the meeting was <B_TIMEX TYPE="DURATION">two hours<E_TIMEX> long
```

“the meeting was two hours in length”

```
the meeting was <B_TIMEX TYPE="DURATION">two hours<E_TIMEX> in length
```

6.8.2 Non-taggable Durations

Generic periods of time, and those without numerals (or designated numeric substitutes), are not taggable:

“the day off”
[no markup for “day”]

“watch TV at night”
[no markup for “night”]

Durations expressed as orders of magnitude are not taggable:

“a matter of days”
[no markup for “days”]

“within hours”
[no markup for “hours”]

6.8.3 Distinguishing DURATION from DATE/TIME

Absolute times and dates are marked with DATE and TIME types even if they are in a phrase which describes a duration.

"twelve twenty to three _p_m"

```
<B_TIMEX TYPE="TIME">twelve twenty<E_TIMEX> to <B_TIMEX TYPE="TIME">three
   _p_m<E_TIMEX>
```

"18th century"

```
<B_TIMEX TYPE="DATE">18th century<E_TIMEX>
```

6.8.4 Distinguishing TIMEX DURATION from NUMEX MEASURE

A thing’s *age*, although related to time, is marked with NUMEX type MEASURE. (See section 7.4.2.1 Standard Measurement Units.)   The line between the *age* of something and the *duration* of something can sometimes be quite arbitrary. The rule of thumb is as follows: If a time expression modifies an adjectival phrases headed with the adjective “old,” the time expression is treated as an age expression (and thus tagged with NUMEX MEASURE). If some other type of adjective is the head, the expression is treated as a time expression (and thus tagged with DURATION).

“eight week old standoff”
```
<B_NUMEX TYPE="MEASURE">eight week<E_NUMEX> old standoff
```

“two month long standoff”
```
<B_TIMEX TYPE="DURATION">two month<E_TIMEX> long standoff
```

7 NUMEX: SPECIFIC GUIDELINES

The NUMEX portion of the task captures a set of useful numeric expressions categorized by the following TYPEs:

MONEY: monetary expression
MEASURE: standard numeric measurement phrases such as age, area, distance, energy, speed, temperature, volume, and weight, plus syntactically-defined measurement phrases
PERCENT: percentage (a fraction expressed in terms of hundredths)
CARDINAL: a numerical count or quantity of some object (in the form of whole numbers, decimals, or fractions)

Note that many of these types could be broken down into subtypes if desired by the end-user. For example, one could envision adding subtypes such as AGE and TEMPERATURE under MEASURE.

7.1 Scope of Numeric Expressions

The word “minus,” or the minus sign, should be included in the tagged numeric expression if it is a negative value.

"minus 15 percent"
```
<B_NUMEX TYPE="PERCENT">minus 15 percent<E_NUMEX>
```

"minus seven degrees"
```
<B_NUMEX TYPE="MEASURE">minus seven degrees<E_NUMEX>
```

7.1.1 Numeric Expressions Plus Units of Measure

As with Hub 4 ’98, the entire string expressing a monetary (MONEY) or percentage (PERCENT) value is to be tagged. This same guideline will apply to the MEASURE TYPE as well, so that the numeric value plus the unit of measure is included within the scope of the tag.

"20 million New Pesos"
```
<B_NUMEX TYPE="MONEY">20 million New Pesos<E_NUMEX>
```

"$42.1 million"
```
<B_NUMEX TYPE="MONEY">$42.1 million<E_NUMEX>
```

"million-dollar conferences"
```
<B_NUMEX TYPE="MONEY">million-dollar<E_NUMEX> conferences
```

"one point four million dollars"
```
<B_NUMEX TYPE="MONEY">one point four million dollars<E_NUMEX>
```

"three dollars and three quarters"
```
<B_NUMEX TYPE="MONEY">three dollars and three quarters<E_NUMEX>
```

"three quarters percent"
```
<B_NUMEX TYPE="PERCENT">three quarters percent<E_NUMEX>
```

"a difference of four percentage points between the two"
```
a difference of <B_NUMEX TYPE="PERCENT">four percentage points<E_NUMEX>
```

```
   between the two
```

“27.5 acres”

```
<B_NUMEX TYPE="MEASURE">27.5 acres<E_NUMEX>
```

“90-degree oven”

```
<B_NUMEX TYPE="MEASURE">90-degree<E_NUMEX> oven
```

If there has been elision of a portion of the expression, the number itself should be tagged:

"some think the rate's as high as twenty percent when it's more like five"

```
some think the rate's as high as <B_NUMEX TYPE="PERCENT">twenty
   percent<E_NUMEX> when it's more like <B_NUMEX
   TYPE="PERCENT">five<E_NUMEX>
```

Combinations of measured quantities should be tagged as a single unit (see section 7.4.2.4 Combinations of Measured Quantities)

"$6 a yard"

```
<B_NUMEX TYPE="MEASURE">$6 a yard<E_NUMEX>
```

[Note that “$6” is not tagged for MONEY because nested expressions are not allowed.]

Dimension words or prepositional phrases following MEASURE expressions should *not* be included within the scope of the tags. For example:

“two miles long”

```
<B_NUMEX TYPE="MEASURE">two miles<E_NUMEX> long
```

“the coast is five miles in length”

```
the coast is <B_NUMEX TYPE="MEASURE">five miles<E_NUMEX> in length
```

“five years old”

```
<B_NUMEX TYPE="MEASURE">five years<E_NUMEX> old
```

“five years of age”

```
<B_NUMEX TYPE="MEASURE">five years<E_NUMEX> of age
```

“a five-foot wide corridor”

```
a <B_NUMEX TYPE="MEASURE">five-foot<E_NUMEX> wide corridor
```

Verbs of measure such as “measure,” and “weigh” are *not* included within the scope of MEASURE expression (just as “cost” is not included within the scope of a MONEY expression).

"it weighed 40 tons"

```
it weighed <B_NUMEX TYPE="MEASURE">40 tons<E_NUMEX>
```

7.1.2 Numeric Expressions without Units of Measure

For the remaining NUMEX type, namely CARDINAL, only the numeral itself is to included within the scope of the tag.

"3 million residents"

```
<B_NUMEX TYPE="CARDINAL">3 million<E_NUMEX> residents
```

7.2 Numeric Expressions Appearing in Succession

Juxtaposed strings expressing values in two different units of measure are to be tagged separately.

“#26 million ($43.6 million)”

```
<B_NUMEX TYPE="MONEY">#26 million<E_NUMEX> (<B_NUMEX TYPE="MONEY">$43.6
   million<E_NUMEX>)
```

“23 degrees Celsius (73 degrees Fahrenheit)”

```
<B_NUMEX TYPE="MEASURE">23 degrees Celsius<E_NUMEX> (<B_NUMEX
   TYPE="MEASURE">73 degrees Fahrenheit<E_NUMEX>)
```

7.3 Approximators and Multipliers in the Modification of Numeric Expressions

7.3.1 Approximators

Modifying words that indicate the approximate value of a number or a “relative position” to a number are generally to be excluded from the NUMEX tag if the modifier indicates only some **minor** imprecision in the known quantity (see also section 7.3.2 Multipliers).

"about 5%"

```
about <B_NUMEX TYPE="PERCENT">5%<E_NUMEX>
```

"over $90,000"

```
over <B_NUMEX TYPE="MONEY">$90,000<E_NUMEX>
```

"65 or older"

```
<B_NUMEX TYPE="MEASURE">65<E_NUMEX> or older
```

NE99 conventions do not allow for the tagging of discontinuous structures. If a modifier occurs in the middle of an otherwise taggable numeric expression, the entire expression should be tagged, regardless of whether the modifier seems to be semantically “approximate” or “indefinite.”

"30 million plus New Pesos"

```
<B_NUMEX TYPE="MONEY">30 million plus New Pesos<E_NUMEX>
```

"three hundred thousand more dollars"

```
<B_NUMEX TYPE="MONEY">three hundred thousand more dollars<E_NUMEX>
```

7.3.2 Multipliers

Modifiers that indicate the multiplied value of a number unit should be included in the tagged string if the modifier is a substitute for a specific digit (or the indefinite article or other quantitative determiner) within the numeric expression.

"several million New Pesos"

```
<B_NUMEX TYPE="MONEY">several million New Pesos<E_NUMEX>
```

"several million dollars"

```
<B_NUMEX TYPE="MONEY">several million dollars<E_NUMEX>
```

```
“a few thousand votes”
a <B_NUMEX TYPE="CARDINAL">few thousand<E_NUMEX> votes
```

In these cases, “several” and “few” are substitutes for some specific digit such as “three” or “four.” Note that the expression remains grammatical if such a digit is substituted for the phrases “a few” or “several” but that the expression “about 10 million New Pesos” does NOT remain grammatical if the approximator “about” is replaced by a digit.

Note that “few,” “couple” etc. are never tagged with the CARDINAL tag -- see section 7.4.4.

Other types of multiplicative cases include actual multiplications such as the following:

"twelve times the national average"

```
<B_NUMEX TYPE="CARDINAL">twelve<E_NUMEX> times the national average
```

7.3.3 Orders of Magnitude

Numeric expressions which give order of magnitude information will be tagged.

"millions of Europeans"

```
<B_NUMEX TYPE="CARDINAL">millions<E_NUMEX> of Europeans
```

"banks are carrying billions of dollars worth"

```
banks are carrying <B_NUMEX TYPE="MONEY">billions of dollars<E_NUMEX> worth
```

7.3.4 The Indefinite Article

The indefinite article can substitute for the digit “one,” in determining whether a string is a markable MONEY, PERCENT, or MEASURE expression.

"a hundred percent"

```
a <B_NUMEX TYPE="PERCENT">hundred percent<E_NUMEX>
```

“a yard of fabric”

```
a <B_NUMEX TYPE="MEASURE">yard<E_NUMEX> of fabric
```

Note that the indefinite article should never be tagged with the CARDINAL tag -- see section 7.4.4.)

7.4 TYPE-Specific Guidelines

The TYPE attribute categorizes the tagged tokens according to measurements of MONEY and other measures (the MEASURE tag) and by numeric form (the PERCENT and CARDINAL tags).

7.4.1 MONEY

Monetary expressions are assigned the TYPE MONEY, unless they are combined with measured quantities (see section 7.4.2.4 Combinations of Measured Quantities.)

"million-dollar conferences"

```
<B_NUMEX TYPE="MONEY">million-dollar<E_NUMEX> conferences
```

"one point four million dollars"

```
<B_NUMEX TYPE="MONEY">one point four million dollars<E_NUMEX>
```

Numeric expressions that do not use currency terms directly to indicate money values are still to be tagged if world knowledge indicates that they are monetary values. The following examples refer to stock prices:

"12 points"

```
<B_NUMEX TYPE="MONEY">12 points<E_NUMEX>
```

"a fixed 106 7/8"

```
a fixed <B_NUMEX TYPE="MONEY">106 7/8<E_NUMEX>
```

7.4.2 MEASURE

The MEASURE type applies to numeric measurement phrases. Taggable expressions include those with standard units of measure, as described in the next section, as well as certain syntactically-defined phrases, discussed in section 7.4.2.2. See also section “7.7 Non-taggable Non-Numeric Expressions” for explicit exclusions.

7.4.2.1 Standard Measurement Units

Taggable MEASURE expressions contain “standard measurement units,” which are defined as measures whose quantity values do not change over time. For example, a “yard” always consists of three feet, so a numeric expression like “ten yards” is taggable. In contrast, in a phrase like “one wave of water after another” the string “one wave of water” is not a MEASURE expression because “wave” does not have a fixed volume. Typical taggable measures are age, area, distance, energy, speed, temperature, volume, and weight. For example:

Age -- the age of a person or thing given in terms of some unit of time, plus the phrase including that unit. In age expressions, the unit of measure is often implied rather than stated.

"i'm 77"

```
i’m <B_NUMEX TYPE="MEASURE">77<E_NUMEX>
```

“she is 30 years old”

```
she is <B_NUMEX TYPE="MEASURE">30 years<E_NUMEX> old
```

"20 and 30 something"

```
<B_NUMEX TYPE="MEASURE">20<E_NUMEX> and <B_NUMEX TYPE="MEASURE">30
   something<E_NUMEX>
```

“a story more than forty years old”

```
a story more than <B_NUMEX TYPE="MEASURE">forty years<E_NUMEX> old
```

Indications of age that are non-numeric are not taggable.

"middle aged"
[no markup]

Area -- the measure of a two-dimensional space.

"27.5 acres"

```
<B_NUMEX TYPE="MEASURE">27.5 acres<E_NUMEX>
```

Distance – the linear measure of the space between two points.

"the lowest legal altitude is 500 feet"

```
the lowest legal altitude is <B_NUMEX TYPE="MEASURE">500 feet<E_NUMEX>
```

"300 feet or below"

```
<B_NUMEX TYPE="MEASURE">300 feet<E_NUMEX> or below
```

“she grew more than 22 inches"

```
she grew more than <B_NUMEX TYPE="MEASURE">22 inches<E_NUMEX>
```

"20 miles from Cairo"

```
<B_NUMEX TYPE="MEASURE">20 miles<E_NUMEX> from <B_ENAMEX
   TYPE="LOCATION">Cairo<E_ENAMEX>
```

“the 30K classical-style race”

```
the <B_NUMEX TYPE="MEASURE">30K<E_NUMEX> classical-style race
```

Energy -- measures of any form of power including electricity, heat, work, radiation, light, or sound.

"60 calories"

```
<B_NUMEX TYPE="MEASURE">60 calories<E_NUMEX>
```

"a powerful earthquake measuring 6.2 on the Richter scale"

```
a powerful earthquake measuring <B_NUMEX TYPE="MEASURE">6.2 on the Richter
   scale<E_NUMEX>
```

Speed -- a measure of the rate of movement usually in units indicating distance over time.

“she drove eighty miles per hour"

```
she drove <B_NUMEX TYPE="MEASURE">eighty miles per hour<E_NUMEX>
```

Temperature -- the degree of hotness or coldness of an object or an environment measured on a standard scale.

"temperatures around zero"

```
temperatures around <B_NUMEX TYPE="MEASURE">zero<E_NUMEX>
```

"a high of 50 degrees Fahrenheit"

```
a high of <B_NUMEX TYPE="MEASURE">50 degrees Fahrenheit<E_NUMEX>
```

Note that “Fahrenheit” and “Celsius” are included within the scope of the tag if they occur.

Volume – a measure of the amount of space occupied by an object or its capacity.

"three cups"

```
<B_NUMEX TYPE="MEASURE">three cups<E_NUMEX>
```

"500,000 cubic yards"

```
<B_NUMEX TYPE="MEASURE">500,000 cubic yards<E_NUMEX>
```

Weight – a measure of the heaviness of an object using any country’s standard measure.

"it weighed 40 tons"

```
it weighed <B_NUMEX TYPE="MEASURE">40 tons<E_NUMEX>
```

"5 pounds of meat"

```
<B_NUMEX TYPE="MEASURE">5 pounds<E_NUMEX> of meat
```

7.4.2.2 Measurement Phrases

Another numeric expression which is marked as type MEASURE are measurement phrases. Such measurement phrases have a distinct syntax, with canonical form that can be represented as follows:

[numeral]-[unit of measure] [head noun]

Note that *hyphenation may not be present*, particularly in speech transcripts. The critical feature of these expressions is that the measurement phrase modifies a head noun. Another key feature of this form is that the noun representing the unit of measure always appears in its *singular* form, despite the plural implication. The “unit of measure” expression can be one of those defined above under “Standard Measurement Units” or can be *any noun* used in this structure, e.g., “a three judge panel.” (There are two exceptions to this -- measures of time duration, which use the TIMEX DURATION tag, and measures of money, which use the NUMEX MONEY tag. See section 7.4.2.5, below.)

The entire measurement phrase, numeral plus unit of measure, is included within the extent of the tag. Dimension words (e.g., “old” of “6-year-old”) and the head noun are *not* included within the scope of the tag. Here are some examples of the canonical form with a range of measurement types:

"their 6-year-old twins"

```
their <B_NUMEX TYPE="MEASURE">6-year<E_NUMEX>-old twins
```

"8,600-square-foot house"

```
<B_NUMEX TYPE="MEASURE">8,600-square-foot<E_NUMEX> house
```

"24-by-24-foot cabin"

```
<B_NUMEX TYPE="MEASURE">24-by-24-foot<E_NUMEX> cabin
```

"a 12-mile traffic jam"

```
a <B_NUMEX TYPE="MEASURE">12-mile<E_NUMEX> traffic jam
```

"a 94-mile-an-hour fastball"

```
a <B_NUMEX TYPE="MEASURE">94-mile-an-hour<E_NUMEX> fastball
```

"90-degree oven"

```
<B_NUMEX TYPE="MEASURE">90-degree<E_NUMEX> oven
```

"112-gallon tank"
```
<B_NUMEX TYPE="MEASURE">112-gallon<E_NUMEX> tank
```

"a three-ton truck"
```
a <B_NUMEX TYPE="MEASURE">three-ton<E_NUMEX> truck
```

"a five-block stretch"
```
a <B_NUMEX TYPE="MEASURE">five-block<E_NUMEX> stretch
```

"six man crew"
```
<B_NUMEX TYPE="MEASURE">six man<E_NUMEX> crew
```

“seven lane highway”
```
<B_NUMEX TYPE="MEASURE">seven lane<E_NUMEX> highway
```

“one column headline”
```
<B_NUMEX TYPE="MEASURE">one column<E_NUMEX> headline
```

"a three-point shot"
```
<B_NUMEX TYPE="MEASURE">three-point<E_NUMEX> shot
```

Note that in the following examples the word “average” is included within the scope of the tag because it acts as the head of the unit of measure:

"a 3.5 grade point average"
```
a <B_NUMEX TYPE="MEASURE">3.5 grade point average<E_NUMEX>
```
[can be paraphrased: “a grade point average of 3.5”]

"a 2.93 earned run average"
```
a <B_NUMEX TYPE="MEASURE">2.93 earned run average<E_NUMEX>
```
[can be paraphrased: “average of the earned runs is 2.93”]

In conjoined expressions, the unit of measure and head noun are typically elided. In such cases, the bare numeral is still tagged as MEASURE.

"a 3 and a 10-gallon tank"
```
a <B_NUMEX TYPE="MEASURE">3<E_NUMEX> and a <B_NUMEX TYPE="MEASURE">10-
   gallon<E_NUMEX> tank
```

7.4.2.3 Some Exceptional Phrases Tagged as MEASURE

Measures in which the numeral itself appears in the plural form are tagged MEASURE if they represent a standard unit of measure, such as age or temperature.

"she’s in her fifties"
```
she’s in her <B_NUMEX TYPE="MEASURE">fifties<E_NUMEX>
```

"high temperatures in the 50’s"
```
high temperatures in the <B_NUMEX TYPE="MEASURE">50’s<E_NUMEX>
```

7.4.2.4   Combinations of Measured Quantities

Some taggable MEASUREs discussed earlier are given in terms of “number per unit,” as in “60 miles per hour” (“60 mph”). Such measure expressions, including those in which “a” is used in place of “per,” are to be tagged as a single MEASURE expression if the unit of measure is “standard” (as defined in 7.4.2.1). Note also that generic time expressions (e.g., “a day”) are allowed in this type of MEASURE expression because these are not measures of time (which are typically not taggable as MEASURE -- see section 7.4.2.5 below); instead, these are measures of quantities *over* time.

"twenty times a day"

```
<B_NUMEX TYPE="MEASURE">twenty times a day<E_NUMEX>
```

[standard measurement unit: “day”]

"$6 a yard"

```
<B_NUMEX TYPE="MEASURE">$6 a yard<E_NUMEX>
```

[standard measurement unit: “yard”]

"about a $1 million a week"

```
about <B_NUMEX TYPE="MEASURE">a $1 million a week<E_NUMEX>
```

[standard measurement unit: “week”]

The numeral can modify any countable noun. The head noun should be included within the scope of the tag:

"thirty-one million flights a year"

```
<B_NUMEX TYPE="MEASURE">thirty-one million flights a year<E_NUMEX>
```

[“flights” is included in the tag]

"missing only five to 10 games per season"

```
missing only <B_NUMEX TYPE="MEASURE">five<E_NUMEX> to <B_NUMEX
TYPE="MEASURE">10 games per season<E_NUMEX>
```

[“games” included in the tag]

Similar, but *not* taggable as MEASURE are expressions like the following, in which the prepositional phrase acts as a time modifier for a preceding clause:

"it will rise 10 percent over the next three years"

```
it will rise <B_NUMEX TYPE="PERCENT">10 percent<E_NUMEX> over the next
   <B_TIMEX TYPE="DURATION">three years<E_TIMEX>
```

"gained two-tenths of 1 percent in December and 1.2 percent in the full year"

```
gained <B_NUMEX TYPE="PERCENT">two-tenths of 1 percent<E_NUMEX> in <B_TIMEX
   TYPE="DATE">December<E_TIMEX> and <B_NUMEX TYPE="PERCENT">1.2
   percent<E_NUMEX> in the full year
```

"happened two times in ’95"

```
happened <B_NUMEX TYPE="CARDINAL">two<E_NUMEX> times in <B_TIMEX
   TYPE="DATE">’95<E_TIMEX>
```

7.4.2.5 Numeric Expressions Not Tagged as MEASURE

Measures of time are not tagged with the NUMEX MEASURE tag (with the exception of age). Instead, they use the TIMEX of type DURATION (see section 6.8)

“it lasted nearly two hours”

```
it lasted nearly <B_TIMEX TYPE="DURATION">two hours<E_TIMEX>
```

"a five-month American campaign"

```
a <B_TIMEX TYPE="DURATION">five-month<E_TIMEX> American campaign
```

Measures of money used in a pre-nominal, modifier role are tagged as NUMEX type MONEY:

“a seven billion dollar tax reduction”

```
a <B_NUMEX TYPE="MONEY">seven billion dollar<E_NUMEX> tax reduction
```

7.4.3 PERCENT

A percentage is a fraction with 100 as the assumed denominator. When tagging a numerical value of percent, the word "percent" or any of its variants is considered part of the number being tagged.

"three quarters percent"

```
<B_NUMEX TYPE="PERCENT">three quarters percent<E_NUMEX>
```

"a difference of four percentage points"

```
a difference of <B_NUMEX TYPE="PERCENT">four percentage points<E_NUMEX>
```

Numeric expressions that do not use percentage terms to indicate percentages are tagged as long as world knowledge indicates that they are expressed in percentages.

"Fees 1 3/4."

```
Fees <B_NUMEX TYPE="PERCENT">1 3/4<E_NUMEX>.
```

"the chance is 50-50"

```
the chance is <B_NUMEX TYPE="PERCENT">50<E_NUMEX>-<B_NUMEX
   TYPE="PERCENT">50<E_NUMEX>
```

7.4.4 CARDINAL

When numerals provide a count or quantity of some object that is not a unit of measurement, they are marked by themselves as CARDINAL. The modified noun is not included within the scope of the tag. This CARDINAL TYPE applies to whole numbers, fractions, and decimals. Indefinite articles, “few,” “couple,” and “several” are *not* tagged as CARDINAL. See also section “7.7 Non-taggable Non-Numeric Expressions” for other explicit exclusions.

"120,000 people"

```
<B_NUMEX TYPE="CARDINAL">120,000<E_NUMEX> people
```

"four houses"

```
<B_NUMEX TYPE="CARDINAL">four<E_NUMEX> houses
```

"24,000 reserve troops"

```
<B_NUMEX TYPE="CARDINAL">24,000<E_NUMEX> reserve troops
```

"lost more than half their value"

```
lost more than <B_NUMEX TYPE="CARDINAL">half<E_NUMEX> their value
```

"reduce the risk of a heart attack upto one-third"

```
reduce the risk of a heart attack upto <B_NUMEX TYPE="CARDINAL">one-
   third<E_NUMEX>
```

"one of every eight people"

```
<B_NUMEX TYPE="CARDINAL">one<E_NUMEX> of every <B_NUMEX
   TYPE="CARDINAL">eight<E_NUMEX> people
```

"about one-third of"

```
about <B_NUMEX TYPE="CARDINAL">one-third<E_NUMEX> of
```

“1.5 times”

```
<B_NUMEX TYPE="CARDINAL">1.5<E_NUMEX> times
```

“hundreds of books”

```
<B_NUMEX TYPE="CARDINAL">hundreds<E_NUMEX> of books
```

If the thing counted is implied rather than explicit, the number is still marked:

"vote was 45 to 3"

```
vote was <B_NUMEX TYPE="CARDINAL">45<E_NUMEX> to <B_NUMEX
   TYPE="CARDINAL">3<E_NUMEX>
```

[implied: “45 votes to 3 votes”]

"a scant 61-59 edge in Parliament"

```
a scant <B_NUMEX TYPE="CARDINAL">61<E_NUMEX>-<B_NUMEX
   TYPE="CARDINAL">59<E_NUMEX> edge in <B_ENAMEX
   TYPE="ORGANIZATION">Parliament<E_ENAMEX>
```

Sports statistics use the CARDINAL tag where appropriate:

"they improved their record to 12-1"

```
they improved their record to <B_NUMEX TYPE="CARDINAL">12<E_NUMEX>-<B_NUMEX
   TYPE="CARDINAL">1<E_NUMEX>
```

[the cardinal is counting wins-losses]

"... to hit .400"

```
... to hit <B_NUMEX TYPE="CARDINAL">.400<E_NUMEX>
```

7.5 Decomposable Idioms

Idioms containing numbers are decomposable. If the requirements for the specific NUMEX TYPE are met, numerals within idioms can be tagged.

"tons of letters"

```
<B_NUMEX TYPE="MEASURE">tons<E_NUMEX> of letters
```

7.6 Non-taggagable Numbers

7.6.1 Non-taggable Numbers within Names

Organization names such as the following are not to be decomposed, i.e., the number should not be marked.

"The 700 Club"
"91 F.M. "
"Channel 5"
"Fox 5"

Numbers in street addresses and street names containing a number should not be tagged as CARDINAL (the address as a whole is tagged as ENAMEX LOCATION).

"4766 Broadway"
"44th Street"
"Highway 1"
"the 38th Parallel"

Products and other types of names such as the following are also not decomposable, even if a number is given.

"20/20"
"Wrangler 4X4"
"one-minute maalox"
"TAGEMET HB 200"

7.6.2 Non-taggable Numbers within Temporal Expressions

Numbers within temporal expressions are not to be tagged separately within the TIMEX expression.

“the 20th century”

```
<B_TIMEX TYPE="TIME">the 20th century<E_TIMEX>
```

"8 p.m."

```
<B_TIMEX TYPE="TIME">8 p.m.<E_TIMEX>
```

"June 1996"

```
<B_TIMEX TYPE="DATE">June 1996<E_TIMEX>
```

7.6.3 Non-taggable Pronoun "One"

When the word “one” is used as a pronoun or has any referential quality, it should not be tagged as a number. For example, the following will go unmarked.

"the only one with effervescent power"
"a pretty ugly one at that"
“one loses track”
“a couple of Miami women, one blond and the other a red head”

When the word “one” is used as a cardinal number, tagging is required.

"It has 1,000 faces. The truth only one."

```
It has <B_NUMEX TYPE="CARDINAL">1,000<E_NUMEX> faces. The truth only
   <B_NUMEX TYPE="CARDINAL">one<E_NUMEX>.
```

"all but one of the plaintiffs"

```
all but <B_NUMEX TYPE="CARDINAL">one<E_NUMEX> of the plaintiffs
```

"There are things worse than war, and one of them is again side"

```
There are things worse than war, and <B_NUMEX TYPE="CARDINAL">one<E_NUMEX>
   of them is again side
```

When the noun being counted is elided, the numeral “one” is still tagged:

“You talked about some of the demands that he made. one that we saw was...”

```
You talked about some of the demands that he made. <B_NUMEX
   TYPE="CARDINAL">one<E_NUMEX> that we saw was...
```

[interpretation is “one [demand] that we saw was...”]

“every county in the state but one is now affected”

```
every county in the state but <B_NUMEX TYPE="CARDINAL">one<E_NUMEX> is now
   affected
```

[interpretation is “…but one [county] is now affected”]

7.6.4 Non-taggable Ordinals

Ordinals are not to be tagged. A NUMEX of type “ordinal” may be included in future evaluations.

"second baseman"
[“second” not tagged]

"we go first to the white house"

```
we go first to the <B_ENAMEX TYPE="ORGANIZATION">white house<E_ENAMEX>
```

[“first” not tagged]

Cardinal numerals used in an “ordinal sense” are also not tagged:

“round two”
[means “second round”; no markup]

“grades four and five”
[means “fourth and fifth grades”; no markup]

7.7 Non-taggable Non-Numeric Expressions

Expressions which do not explicitly contain numerals are generally not taggable as either MEASURE or CARDINAL. For example, the following do not qualify as numerals in MEASURE or CARDINAL expressions:

- the adverbs “once” “twice” etc.
- “single,” “double,” “quadruple” etc.
- “both”
- vague quantifiers like “many,” “lots,” “multiple” etc.

Thus, these strings are not tagged:

"it happened once before"
[no markup]

“twice a week”
[no markup]

"single trip"
[no markup]

"triple toe loop"
[no markup]

"speaks both languages"
[no markup]

“a few steps”
[no markup]

“many times a day”
[no markup]

“a few cases of harassment”
[no markup]

Vague references to plurals are not to be tagged:

"a number of slaves"
[no markup]

“multiple felonies”
[no markup]

For MONEY, PERCENT, and MEASURE expressions, the indefinite article *can* often substitute for the numeral “one” (see section 7.3.3). Thus, the following is markable.

“over a foot of rain”

```
over <B_NUMEX TYPE="MEASURE">a foot<E_NUMEX> of rain
```

APPENDIX A. TOKENIZATION RULES

Text elements annotated for NE99 task must consist of one or more complete tokens. This document explains where boundaries of tagged strings are meant to be located when there is NO explicit whitespace between alphanumeric characters and a punctuation mark or other special character.

A.1 Newswire Articles and Other Originally Written Text

A.1.1 Punctuation Marks and Special Characters

Punctuation marks and special characters are normally considered separate tokens in newswire articles.

A.1.1.1 Examples with Period

Examples with periods may include both sentence-end punctuation and ellipsis.

"...Jaguar company in Britain."

```
...<B_ENAMEX TYPE="ORGANIZATION">Jaguar<E_ENAMEX> company in <B_ENAMEX
   TYPE="LOCATION">Britain<E_ENAMEX>.
```

A.1.1.2 Examples with Hyphen or Dash

"Chicago-based"

```
<B_ENAMEX TYPE="LOCATION">Chicago<E_ENAMEX>-based
```

"U.S.-based"

```
<B_ENAMEX TYPE="LOCATION">U.S.<E_ENAMEX>-based
```

"U.S.-Japan trade negotiations"

```
<B_ENAMEX TYPE="LOCATION">U.S.<E_ENAMEX>-<B_ENAMEX
   TYPE="LOCATION">Japan<E_ENAMEX> trade negotiations
```

"an Eaton-Sumitomo joint venture"

```
an <B_ENAMEX TYPE="ORGANIZATION">Eaton<E_ENAMEX>-<B_ENAMEX
   TYPE="ORGANIZATION">Sumitomo<E_ENAMEX> joint venture
```

"PHILADELPHIA--A new recycling center has been built."

```
<B_ENAMEX TYPE="LOCATION">PHILADELPHIA<E_ENAMEX>--A new recycling center
   has been built.
```

A.1.1.3 Examples with Apostrophe

Examples with apostrophe primarily include possessives.

"California's"

```
<B_ENAMEX TYPE="LOCATION">California<E_ENAMEX>'s
```

"Guiness' Schenley Industries"

```
<B_ENAMEX TYPE="ORGANIZATION">Guiness<E_ENAMEX>' <B_ENAMEX
   TYPE="ORGANIZATION">Schenley Industries<E_ENAMEX>
```

A.1.1.4 Examples with Parentheses or Quotes

"(IBM)"

(<B_ENAMEX TYPE="ORGANIZATION">IBM<E_ENAMEX>)

""IBM stock fell today," he said" [note the double quote preceding IBM]

```
"<B_ENAMEX TYPE="ORGANIZATION">IBM<E_ENAMEX> stock fell today," he said
```

A.1.2 Internal Punctuation mark or other Special Character

When a proper name or number contains an internal punctuation mark or other special character, the word containing that character is treated as just one token.

A.1.2.1 Examples with Period

A period that marks an abbreviation is considered part of the abbreviation token, even when the abbreviation appears at the end of a sentence.

"U.K. industry"

```
<B_ENAMEX TYPE="LOCATION">U.K.<E_ENAMEX> industry
```

"Microtest Inc."

```
<B_ENAMEX TYPE="ORGANIZATION">Microtest Inc.<E_ENAMEX>
```

"Spokane, Wash."

```
<B_ENAMEX TYPE="LOCATION">Spokane<E_ENAMEX>, <B_ENAMEX
   TYPE="LOCATION">Wash.<E_ENAMEX>
```

"Limousines are manufactured in the U.K."

```
 Limousines are manufactured in the <B_ENAMEX
   TYPE="LOCATION">U.K.<E_ENAMEX>
```

A period used as a decimal marker is considered integral to the number token.

"$5.10"

```
<B_NUMEX TYPE="MONEY">$5.10<E_NUMEX>
```

A.1.2.2 Examples with Hyphen or Dash

(See also section A.1.3, “Hyphen at End of Line”)

"F. Gregory Fitz-Gerald"

```
<B_ENAMEX TYPE="PERSON">F. Gregory Fitz-Gerald<E_ENAMEX>
```

"Prudential-Bache Securities"

```
<B_ENAMEX TYPE="ORGANIZATION">Prudential-Bache Securities<E_ENAMEX>
```

"one-hundred percent"

```
<B_NUMEX TYPE="PERCENT">one-hundred percent<E_NUMEX>
```

A.1.2.3 Examples with Slash

"The venture will be called Quality Spring/Togo Inc."

```
The venture will be called <B_ENAMEX TYPE="ORGANIZATION">Quality Spring/
   Togo Inc.<E_ENAMEX>
```

"10/13/89"

```
<B_TIMEX TYPE="DATE">10/13/89<E_TIMEX>
```

A.1.2.4 Examples with Other Punctuation

These examples of other punctuation include special uses of apostrophes.

"McDonald's burger company"

```
<B_ENAMEX TYPE="ORGANIZATION">McDonald's<E_ENAMEX> burger company
```

"back in '87"

```
back in <B_TIMEX TYPE="DATE">'87<E_TIMEX>
```

A.1.2.5 Examples with Special Characters

"S&P 500 Index"

```
<B_ENAMEX TYPE="ORGANIZATION">S&P<E_ENAMEX> 500 Index
```

A.1.3 Hyphen at End of Line

When a hyphen is used at the end of a line to separate a single word into two parts, the word is treated as a single token.

"Phila-
delphia"

```
<B_ENAMEX TYPE="LOCATION">Phila-
delphia<E_ENAMEX>
```

If, however, the word is naturally hyphenated and the hyphenated word just happens to be broken at the hyphen at the end of a line, the parts of the word are treated as separate tokens.

"Chicago-
based"

```
<B_ENAMEX TYPE="LOCATION">Chicago<E_ENAMEX>-
based
```

A.2 Speech Transcriptions

The rules in this section are instructions for the benefit of human annotators. Punctuation marks are often seen in human transcription, although they will be stripped in snorification and will not be present in speech recognizer transcripts. The transcription conventions for human transcription state that all the following punctuation may be found in transcriptions. Some of the following examples include capitalized letters, because the transcription guidelines state that they will be used in human transcription; however, experience has shown that this is not usually the case, and when capitalization **is** found in transcriptions it is often random or unreliable.

A.2.1 Punctuation Marks and Special Characters

Punctuation marks and special characters are normally considered separate tokens in speech transcription.

A.2.1.1 Examples with Period

Terminal periods may be seen in speech transcriptions, but ellipsis will not.

"this commitment will allow unscom to fulfill its mission."

```
this commitment will allow <B_ENAMEX TYPE="ORGANIZATION">unscom<E_ENAMEX>
   to fulfill its mission.
```

A.2.1.2 Examples with Apostrophe

Apostrophes in transcribed speech will be found in the case of possessive constructions.

"at the end of president bush’s administration"

```
at the end of president <B_ENAMEX TYPE="PERSON">bush<E_ENAMEX>’s
   administration
```

"in addition to the riadys’ donations"

```
in addition to the <B_ENAMEX TYPE="PERSON">riadys<E_ENAMEX>’ donations
```

A.2.1.3 Examples with Other Punctuation

Only the following additional punctuation marks will be included for ease of (human) reading.

Question marks (?) should be added at the end of interrogative sentences. Commas (,) should be added between clauses.

A.2.2 Internal Punctuation Mark or Other Special Character

When a proper name or number contains an internal punctuation mark or other special character, the word containing that character is treated as just one token.

A.2.2.1 Examples with Underscores

Acronyms or single initials will use underscores as delimiters. All associated underscores must be enclosed in the tag, including the leading underscore.

```
"_c_n_n"
<B_ENAMEX TYPE="ORGANIZATION">_c_n_n<E_ENAMEX>

"_a_t and _t"
<B_ENAMEX TYPE="ORGANIZATION">_a_t and _t<E_ENAMEX>
```

Abbreviations may remain as abbreviations when they are used as part of a title. Otherwise they will be spelled out.

"Today I went to see Dr. Brown"

"Today I went to see the doctor"

APPENDIX B. SPECIAL FEATURES FOUND ONLY IN TRANSCRIBED SPEECH

The output from human transcriptions and speech recognition systems will differ. Human transcriptions will be marked up by humans, then snorified and normalized to produce the human-created scoring keys. Speech recognizer output will be marked up by machines, normalized, then scored against keys.

B.1 Disfluencies

In the cases of both corrections and partial repeats, a hyphen was previously inserted to show the interruption of the incomplete word. This use of the hyphen has now been replaced by an SGML tag indicating a fragment. The tag will not be visible to human annotators.

B.1.1 Repetition of Complete Words

Tag all instances of entities even when they are repeated either for emphasis or correction.

"they're from illinois small town illinois"

```
they're from <B_ENAMEX TYPE="LOCATION">illinois<E_ENAMEX> small town
   <B_ENAMEX TYPE="LOCATION">illinois<E_ENAMEX>
```

"think +dole dole is tough at all"

```
think <B_ENAMEX TYPE="PERSON">+dole<E_ENAMEX> <B_ENAMEX
   TYPE="PERSON">dole<E_ENAMEX> is tough at all
```

"lebanon lebanon lebanon lebanon"

```
<B_ENAMEX TYPE="LOCATION">lebanon<E_ENAMEX> <B_ENAMEX
   TYPE="LOCATION">lebanon<E_ENAMEX> <B_ENAMEX
   TYPE="LOCATION">lebanon<E_ENAMEX> <B_ENAMEX
   TYPE="LOCATION">lebanon<E_ENAMEX>
```

"frank contreras {breath frank contreras"

```
<B_ENAMEX TYPE="PERSON">frank contreras<E_ENAMEX> {breath <B_ENAMEX
   TYPE="PERSON">frank contreras<E_ENAMEX>
```

B.1.2 Partial Repetition

If a fragment of a word or entity name occurs at either the beginning or the end of a complete entity name, the fragment will be left out of the tagged name.

"at sev seven this evening"

```
at sev<FRAGMENT> <B_TIMEX TYPE="TIME">seven this evening<E_TIMEX>
```

"it costs ninety dollars dollars to"

```
it costs <B_NUMEX TYPE="MONEY">ninety dollars<E_NUMEX> dollars to
```

If, however, the fragment occurs within the bounds of the minimal string identifiable as the entity name intended by the speaker, the fragment will be included in the string.

"seven fi fifty nine"

```
<B_TIMEX TYPE="TIME">seven fi<FRAGMENT> fifty nine<E_TIMEX>
```

B.1.3 Corrections

If a correction does not fall within a valid Named Entity expression, it will not be tagged.

"be sure they credit clear clinton policies"

```
be sure they credit clear<FRAGMENT> <B_ENAMEX
```

```
   TYPE="PERSON">clinton<E_ENAMEX> policies
```

[Speaker corrected himself. "clear" was broken off <fragment>; sentence was intended to be "be sure they credit clinton policies." Absence of visible hyphen at break point may make this more difficult to recognize.]

If however, a correction falls within or comprises a valid Named Entity expression, it will be tagged even when the annotator can determine that the words were said in error.

"at five twenty i mean five thirty"

```
at <B_TIMEX TYPE="TIME">five twenty<E_TIMEX> i mean <B_TIMEX
   TYPE="TIME">five thirty<E_TIMEX>
```

"richard i mean john allen"

```
<B_ENAMEX TYPE="PERSON">richard<E_ENAMEX> i mean <B_ENAMEX
   TYPE="PERSON">john allen<E_ENAMEX>
```

B.2 Interjections

The following interjections are listed in the transcription guidelines for NE markup, and will be considered as words for English language transcription:

"%um, %uh, %oh, %uhoh, %mhm, %uhhuh, %okay, %whoa, %whew, %yeah, %jeeze"

Interjections will never be marked by themselves, but they will be included in the tagged string if they occur within the bounds of a taggable Named Entity (see section 4.3, "Effects of Tokenization Conventions").

"richard allen %uh men's society"

```
<B_ENAMEX TYPE="ORGANIZATION">richard allen %uh men's society<E_ENAMEX>
```

"in %uh nineteen %uh ninety four"

```
in %uh <B_TIMEX TYPE="DATE">nineteen %uh ninety four<E_TIMEX>
```

B.3 Overlap

Overlap occurs when two or more speakers speak at the same time. This will be indicated in output from human transcribers by an SGML "overlap" tag. This SGML tag should not preclude insertion of NE tags (see section B.4.2, "SGML Tags") .

B.4 Shortrefs and SGML Markup by Transcribers

The output from human transcribers will include many symbols and short-refs, in addition to more normal SGML tags. The following list is intended to be as complete a list of expected symbols and tags as we can provide in this document.

B.4.1 Shortrefs

The following symbols will be visible to annotators. These symbols are to be included in the NE tag when they are adjacent to a Named Entity (see also section B.1.1, Repetition of Complete Words).

+	mispronounced - the transcription reflects what was believed to be the speaker's intent.
@	misspelled, uncertain spelling.
_	acronym.
^	proper name.
*	idiosyncratic speech, made-up words.

Noises made by speaker; only the listed notations are allowed:

{breath, {laugh, {cough, {sneeze, {lipsmack

When these notations are located within the bounds of a taggable string, they will be included. When they fall outside a taggable string, they will not be tagged.

"frank contreras {breath frank contreras"

```
<B_ENAMEX TYPE="PERSON">frank contreras<E_ENAMEX> {breath <B_ENAMEX
   TYPE="PERSON">frank contreras<E_ENAMEX>
```

["{breath" falls outside of person name, not tagged]

"in the wall street {cough journal today"

```
in the <B_ENAMEX TYPE="ORGANIZATION">wall street {cough journal<E_ENAMEX>
   today
```

["{cough" falls inside newspaper name, included in tag]

If, however, a shortref symbol is placed outside of the string, but in direct contact with part of the tagged string (no whitespace between them), the shortref must be included in the tagged string.

"this is @glen bunting"

```
this is <B_ENAMEX TYPE="PERSON">@glen bunting<E_ENAMEX>
```

B.4.2 SGML Tags

The following identifiers will be found in speech transcriptions, although they will not be visible to annotators using the Alembic Workbench tool for markup.

BACKGROUND COMMENT CONTRACTION
FRAGMENT HYPHEN SECTION
SEGMENT SYNC TIME
TURN UTF BN_EPISODE_TRANS

B_UNCLEAR E_UNCLEAR
B_OVERLAP E_OVERLAP
B_NOSCORE E_NOSCORE
B_FOREIGN E_FOREIGN
B_ASIDE E_ASIDE

B_ENAMEX E_ENAMEX
B_TIMEX E_TIMEX
B_ NUMEX E_NUMEX

For instance, overlap will now appear this way:

"in the _l_a
times from"

```
in the _l_a
<B_OVERLAP STARTTIME="1724.051312" ENDTIME="1725.180813">
times from
<E_OVERLAP>
```

B.5 Errors in Transcription

Expressions should be tagged regardless of whether transcribed with conventional spelling. Misspellings and truncations should be tagged. All types of spelling variations will be normalized before scoring.

B.5.1 Misspellings, Truncations, and Unconventional Treatments

"_n double _a_c_p"

```
<B_ENAMEX TYPE="ORGANIZATION">_n double _a_c_p<E_ENAMEX>
```

[conventional spelling would be _n_a_a_c_p]

"mid autumm"

```
mid <B_TIMEX TYPE="DATE">autumm<E_TIMEX>
```

["autumn" misspelled]

"in the _s_n_p index today"

```
in the <B_ENAMEX TYPE="ORGANIZATION">_s_n_p<E_ENAMEX> index today
```

["_s_n_p" is a mis-transcription of "_s and _p"]

"from cairo _n_p sunni khalid reports"

```
from <B_ENAMEX TYPE="LOCATION">cairo<E_ENAMEX> <B_ENAMEX
   TYPE="ORGANIZATION" STATUS="OPT">_n_p<E_ENAMEX> <B_ENAMEX
   TYPE="PERSON">sunni khalid<E_ENAMEX> reports
```

["_n_p" is assumed to be a mis-transcription of "_n_p_r's," but lack of certainty dictates the OPT status.]

"in her 1950's"

```
in her <B_NUMEX TYPE="MEASURE">1950's<E_NUMEX>
```

["in her 1950's" is assumed to be a mis-transcription of "in her 50's," so this is marked as MEASURE instead of DATE. Lack of certainty regarding transcription may dictate using the OPT status.]

The above examples also appear in newswire articles when errors in capitalization occur and they are markable.

B.5.2 True Spelling Variants

True spelling variants are words that have alternate spellings, but each is correct. A standard dictionary has been used in the past as the authoritative reference for spelling variants, and can be referred to if necessary. (See section 2.2 for the Authoritative Reference Materials.)

APPENDIX C. GUIDELINES FOR MARKUP OF NEWSWIRE DATA

Both broadcast news and newswire data will be used in this year’s Named Entity task. This appendix identifies features unique to the newswire data which affect its annotation.

C.1 Data to be Annotated

Unlike broadcast news data, newswire data contains sections which should not be marked.

C.1.1 Taggble Text

All text within the following newswire tags should be marked for Named Entity:

```
<HEADLINE> ... </HEADLINE>
<TEXT> ... </TEXT>
```

C.1.2 Non-Taggable Text

- Any text within the following newswire tags is not markable:

```
<HEADER> ... </HEADER>
<BODY> ... </BODY>
<SLUG> ... </SLUG>
<TRAILER> ... </TRAILER>
```

- Any text outside of newswire SGML tags is not markable.
- Any text in tabular format (or what is meant to be tabular in final output) is not markable.

C.2 Tokenization Rules

C.2.1 Character Codes

The electronic versions of some newswire data contain non-standard character codes substituted for the standard typographical symbols that would appear in the final printed copy of the newspaper.

Special characters and character codes beginning with ampersands will be marked if they occur within a markable string.

“a partner in the McGlashan &AMP; Sarrail firm in San Mateo”

```
a partner in the <B_ENAMEX TYPE="ORGANIZATION">McGlashan &AMP;
   Sarrail<E_ENAMEX> firm in San Mateo
```

“1{ tablespoons unsalted butter”

```
<B_NUMEX TYPE="MEASURE">1{ tablespoons<E_NUMEX> unsalted butter
```

Some character codes appear on the edges of markable strings. If these special characters are normally markable when they occur in their normal form, they will be included within the scope of the tag. For example, dollar signs ($) are included within the scope of MONEY expressions, so “dlrs” in the following example is also included:

“The IMF last year withheld a dlrs 20 million loan”

```
...a <B_NUMEX TYPE=”MONEY”>dlrs 20 million<E_NUMEX> loan
```

However, numerals that have been substituted with character codes are only optionally included. The reference key will make use of the ALT attribute to mark these:

“} cup sherry vinegar”

```
<B_NUMEX TYPE="MEASURE" ALT="cup">} cup<E_NUMEX> sherry vinegar
```

“about { hour”

```
about <B_TIMEX TYPE="DURATION" ALT="hour">{ hour<E_TIMEX>
```